UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ

***CAMPUS* DE CASCAVEL**

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AGRÍCOLA

MODELAGEM DO COEFICIENTE DE SORÇÃO DO SOLO DE POLUENTES ORGÂNICOS PERSISTENTES NO MEIO AMBIENTE

CARLOS JOSÉ MARIA OLGUIN

CASCAVEL – Paraná – Brasil

FEVEREIRO 2017

CARLOS JOSÉ MARIA OLGUIN

MODELAGEM DO COEFICIENTE DE SORÇÃO DO SOLO DE POLUENTES ORGÂNICOS PERSISTENTES NO MEIO AMBIENTE

Tese apresentada ao Programa de Pós Graduação em Engenharia Agrícola em cumprimento aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Recursos Hídricos e Saneamento Ambiental.

Orientador: Prof. Dr. Silvio César Sampaio
Co-orientador: Prof. Dr. Ralpho Rinaldo dos Reis

CASCAVEL – Paraná – Brasil

FEVEREIRO 2017

Dados Internacionais de Catalogação-na-Publicação (CIP)
Ficha catalográfica elaborada por Rosângela A. A. Silva – CRB 9ª/1810

O38m Olguin, Carlos José Maria.
Modelagem do coeficiente de sorção do solo de poluentes orgânicos persistentes no meio ambiente. /Carlos José Maria Olguin. — Cascavel - PR: UNIOESTE, 2017. — xiii, 115 p.

Orientador: Prof. Dr. Silvio Cesár Sampaio

Co-orientador: Prof. Dr. Ralpho Rinaldo dos Reis

Tese (Doutorado) – Universidade Estadual do Oeste do Paraná, Campus de Cascavel, 2017
Programa de Pós-Graduação em Engenharia Agrícola, Centro de Ciências Exatas e Tecnológicas.

Bibliografia

1. Avaliação de riscos ambientais. 2. Coeficiente de partição. 3. Modelos QSPR. I. Universidade Estadual do Oeste do Oeste Paraná. II. Titulo

CDD 20.ed. 631.41

As revisões de português, de inglês e de normas de monografias do PGEAGRI foram feitas pelos Prof. Dr. José Carlos da Costa em 20/04/2017.

CARLOS JOSÉ MARIA OLGUIN

Modelagem do Coeficiente de Sorção do Solo de Poluentes Orgânicos Persistentes

Tese apresentada ao Programa de Pós-Graduação em Engenharia Agrícola em cumprimento parcial aos requisitos para obtenção do título de Doutor em Engenharia Agrícola, área de concentração Recursos Hídricos e Saneamento Ambiental, linha de pesquisa Recursos Hídricos, APROVADO(A) pela seguinte banca examinadora:

Orientador(a) - Silvio César Sampaio
Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Marcio Antonio Vilas Boas
Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Ralpho Rinaldo dos Reis
Universidade Estadual do Oeste do Paraná - Campus de Cascavel (UNIOESTE)

Jonathan Dieter
Universidade Federal do Paraná - Campus de Palotina (UFPR)

Jiam Pires Frigo
Universidade Federal da Integração Latino-Americana (Unila)

Cascavel, 17 de fevereiro de 2017

BIOGRAFIA

Carlos José Maria Olguin é natural de Salta, Pcia. de Salta, Argentina. Cursou Licenciatura de Sistemas na Universidad de Belgrano – UB, Buenos Aires, Argentina (1985). Em 1990 obteve seu título de Mestre em Engenharia Elétrica (área de concentração: Automação) pela Universidade Estadual de Campinas – UNICAMP, Campinas-SP, Brasil. Desde 2013 realiza estudos de doutorado no programa de Pós-graduação em Engenharia Agrícola, na área de concentração de Recursos Hídricos e Saneamento Ambiental, sob orientação do Prof. Dr. Silvio Cesar Sampaio e coorientação do Prof. Dr. Ralpho Rinaldo dos Reis. Desde o início da sua vida profissional, em 1980, trabalhou em empresas do setor privado, autarquias públicas e centros de pesquisa da área de Tecnologia da Informação (TI), tanto no Brasil quanto na Argentina. Em 1997 ingressou, mediante concurso público, na carreira de professor universitário no Estado do Paraná atuando como professor assistente no curso de Ciência da Computação da Universidade Estadual de Maringá – UEM até 2005, ano em que se transferiu para a Universidade Estadual do Oeste do Paraná – UNIOESTE, onde atua como professor no curso de Ciência da Computação do *campus* de Cascavel.

“...Dê-me, Senhor, agudeza para entender, capacidade para reter,
método e faculdade para aprender, sutileza para interpretar,
graça e abundância para falar, acerto ao começar,
direção ao progredir e perfeição ao concluir...”

São Tomás de Aquino

Para a mulher que rio, choro, canto, amo, vivo ...
Com todo o meu amor, dedico a minha esposa
***Conceição de Fátima Alves Olguin**.*

AGRADECIMENTOS

À minha família, em especial, a minha esposa Conceição de Fátima Alves Olguin e aos meus filhos Mariana, Gabriela e Enrique, pela paciência, pelo suporte e por compreender que o tempo que deixamos de ficar juntos devido aos meus estudos foi necessário; aos meus pais Alfredo Eduardo Olguin e Susana Angélica Lona, exemplos máximos de honestidade e bondade; aos meus irmãos Alfredo, Ignácio, Paola e Marcelo; aos meus sogros Antônio e Irene; aos meus cunhados e cunhadas Carlos, Horácio, Tânea e Elídio, José Roberto e Josiane, Marcos e Inês e Robson e Patrícia, por terem contribuído, saibam ou não, cada um do seu jeito, na concretização deste trabalho, e por tornarem minha vida feliz.

Aos professores Silvio César Sampaio e Ralpho Rinaldo dos Reis, orientadores e amigos, por todo conhecimento repassado, pelas conversas inspiradoras, pela motivação necessária nas horas complicadas e pelas contribuições durante toda a pesquisa.

Aos meus amigos/irmãos Carlos Maria Punta Raffo, Antônio Francisco Grando, José Grando e Anibal Mantovani Diniz.

À banca avaliadora, pela leitura e contribuições para melhoria do trabalho, em especial ao Prof. Dr. Jiam Pires Frigo, pela amizade e apoio no início desta jornada.

Ao Programa de Pós Graduação em Engenharia Agrícola, pela oportunidade e suporte durante todo o curso.

A todos os professores do Programa, especialmente àqueles que ministraram aula para mim, em ordem cronológica, Miguel Angel Uribe Opazo, Eloy Lemos de Mello, Erivelto Mercante, Marcio Antonio Vilas Boas, Benedito Martins Gomes, Mariângela Alice Pieruccini, Luciana Pagliosa Carvalho Guedes, Jerry Adriani Johan, Silvia Renata Machado Coelho e Divair Christ, quer seja como aluno ouvinte ou especial nos primeiros passos dados para ingressar no Programa ou já como aluno regular do Programa, pelo aprendizado recebido.

A todos aqueles que, de alguma forma, contribuíram para a realização deste trabalho, de suma importância para minha vida profissional e pessoal, em especial, a Marcelo Remor, aos “vizinhos” e a todos os colegas do grupo de pesquisa.

MODELAGEM DO COEFICIENTE DE SORÇÃO DO SOLO DE POLUENTES ORGÂNICOS PERSISTENTES NO MEIO AMBIENTE

RESUMO

O coeficiente de sorção do solo normalizado para o conteúdo de carbono orgânico (Koc) é um parâmetro físico-químico utilizado em avaliações de risco ambiental e na determinação do destino final das substâncias químicas lançadas na natureza. Vários modelos para prever este parâmetro foram propostos com base na relação entre LogKoc e LogP. A dificuldade e o custo para a obtenção de valores experimentais de LogP levaram ao desenvolvimento de algoritmos para calculá-los. Assim, no primeiro artigo desta tese foram considerados diversos algoritmos gratuitos para cálculo de LogP, e concluiu-se que os melhores modelos QSPR para predizer o coeficiente de sorção do solo de compostos orgânicos não iónicos foram obtidos usando os algoritmos ALOGPs, KOWWIN e XLOGP3. Neste estudo, foram demonstradas a importância e a utilidade do teste de equivalência estatística utilizado, dados que nos permitiram afirmar que os modelos obtidos dos algoritmos considerados são estatisticamente equivalentes. Assim, na impossibilidade de obterem-se valores de LogP a partir de um dos algoritmos, valores obtidos por outro podem ser usados. Verificou-se ainda que os modelos apresentados neste estudo possuem qualidade estatística e capacidade de predição compatíveis à de modelos mais complexos, publicados recentemente na área. Adicionalmente, a necessidade de se realizar a validação da predição de um modelo QSPR a partir de um conjunto de dados que não foi utilizado na geração do modelo é uma prática bem aceita na área. Nesse contexto, alguns trabalhos exploraram o impacto que diversos tamanhos de conjuntos de treinamento teriam na capacidade de predição dos modelos QSPR gerados, não chegando a resultados conclusivos. Assim, no segundo artigo desta tese, foi mostrado que, a partir de conjuntos de treinamento não tão grandes, modelos QSPR estatisticamente equivalentes podem ser desenvolvidos e que tais modelos têm capacidade de predição similar daqueles criados a partir de um conjunto de treinamento maior. Para isto, modelos foram gerados considerando valores de LogP do conjunto de treinamento total, calculados com o algoritmo ALOGPs e também com subconjuntos do mesmo (i.e., metades, quartos e oitavos). Este estudo, assim como o anterior, confirmou a importância do uso do teste de equivalência estatística utilizado nesta tese já que foi verificado que, seguindo os procedimentos adotados, os modelos obtidos com subconjuntos do conjunto de treinamento são estatisticamente equivalentes.

Palavras-chave: risco ambiental, coeficiente de partição, modelos QSPR.

MODELING OF SOIL SORPTION COEFFICIENT FROM PERSISTENT ORGANIC POLLUTANTS IN THE ENVIRONMENT

ABSTRACT

The soil sorption coefficient normalized for organic carbon content (Koc) is a physicochemical parameter used in environmental risk assessments to determine the final destination of chemicals released in the environment. So, in oreder to predict this parameter, several models were proposed based on the relationship between LogKoc and LogP. The difficulty and cost to obtain experimental values of LogP have drawn to the algorithms development to calculate those values. Thus, in the first paper of this thesis, several free algorithms were considered to calculate LogP, and it was concluded that the best QSPR models to predict soil sorption coefficient of organic nonionic compounds were obtained using ALOGPs, KOWWIN and XLOGP3 algorithms. This study demonstrated the importance and usefulness of the statistical equivalence test used, since it allowed us to state that the models obtained from the considered algorithms are statistically equivalent. In this study, the both importance and usefulness of the statistical equivalence test were proved. These data allowed us to state that the models that have been obtained from the algorithms are statistically equivalent. Thus, in the impossibility of obtaining LogP values based on one of the algorithms, values obtained by another one of them can be used. It was also observed that the models presented in this study presented statistical quality and predictive capacity compatible with more complex models recently published in the area. In addition, it is a well accepted practice in the area the requirement to validate the prediction of a QSPR model from a data set that was not used in the model generation. In this context, some studies have explored the impact that several sizes of training sets would have on the predictive capacity of the generated QSPR models, consequently not reaching conclusive results. Thus, the second paper has been shown that, from not so large training sets, statistically equivalent QSPR models can be developed and that these models have similar predictive capacity to those ones created from a larger training set. Therefore, models were generated considering LogP values of the total training set, calculated with the ALOGPs algorithm and also with subsets of itself (i.e., halves, quarters and eighths). This study, just like the previous one, has confirmed the importance of using the statistical equivalence test since it was ascertained that, following the adopted procedures, the models obtained with subsets of the training set are statistically equivalent.

Keywords: environmental risk, partition coefficient, QSPR models.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS – PARTE GERAL

ARTIGO 1

ARTIGO 2

LISTA DE ABREVIATURAS E SÍMBOLOS

CCC_{cv}	Coeficiente de concordância da correlação da validação cruzada
CCC_{ext}	Coeficiente de concordância da correlação do conjunto de teste
CCC_{tr}	Coeficiente de concordância da correlação do conjunto de treinamento
F	Estatística F
Koc	Coeficiente de sorção no solo normalizado para o conteúdo de carbono orgânico (L kg^{-1})
MSE	Erro quadrado médio
P	Coeficiente de partição octanol/água
$PRESS_{cv}$	Soma dos quadrados dos resíduos da validação cruzada
$PRESS_{ext}$	Soma dos quadrados dos resíduos do conjunto de teste
Q^2_{LOO}	Coeficiente de determinação da validação LOO (*Leave-One-Out*)
Q^2_{LMO}	Coeficiente de determinação da validação LMO (*Leave-Many-Out*)
Q^2_{Yscr}	Coeficiente de determinação do teste Y-*scrambling*
R^2	Coeficiente de determinação
R^2_{adj}	Coeficiente de determinação ajustado
R^2_{ext}	Coeficiente de determinação do conjunto de teste
R^2_{Yscr}	Coeficiente de determinação do teste Y-*scrambling*
$\overline{r^2_m}$	Critério de Roy: média
Δr^2_m	Critério de Roy: delta
$RMSE\_AV_{Yscr}$	Raiz do erro quadrado médio do teste Y-*scrambling*
$RMSE_{cv}$	Raiz do erro quadrado médio da validação cruzada
$RMSE_{ext}$	Raiz do erro quadrado médio do conjunto de teste
$RMSE_{tr}$	Raiz do erro quadrado médio do conjunto de treinamento
RSS_{tr}	Soma dos quadrados dos resíduos do conjunto de treinamento

1 INTRODUÇÃO

Os agrotóxicos são usados na agricultura para o controle de pragas, ervas daninhas e doenças nas plantas. A sua aplicação é ainda o meio mais eficaz e aceito para proteger as plantas de parasitas e tem contribuído para aumentar, significativamente, a produtividade de diversas culturas e, consequentemente, a produtividade agrícola (BOLOGNESI, 2003).

No entanto, o uso desses produtos para o tratamento de plantas pode apresentar riscos toxicológicos e ecotoxicológicos como, por exemplo, a exposição de organismos não-alvo e a ocorrência de efeitos colaterais indesejáveis em algumas espécies, comunidades ou mesmo em ecossistemas como um todo (JURASKE et al., 2007).

Considerando-se que, em alguns casos, menos de 0,1% da quantidade de pesticidas aplicados alcançam o alvo e o restante tem potencial para se mover para outros compartimentos ambientais, como águas superficiais e subterrâneas (ARIAS-ESTÉVEZ et al., 2008), pode-se inferir que as águas subterrâneas podem ser contaminadas pela percolação da água mediante a lixiviação de pesticidas e outros compostos químicos presentes no solo.

Um fator que pode ser utilizado para estimar a probabilidade de um determinado composto atingir o lençol freático é o coeficiente de sorção no solo desse composto, isto é, quanto menor for o coeficiente de sorção do composto, maior será seu potencial de lixiviação, aumentando a probabilidade de essa substância contaminar as águas subterrâneas (REIS; SAMPAIO; MELO, 2013).

Considerando que o Brasil é o maior consumidor de agrotóxicos do mundo e, ao mesmo tempo, possuidor de uma das maiores reservas de água, estudos que permitam avaliar a possibilidade de riscos ambientais provocados pelo uso de compostos químicos são

de fundamental importância para órgãos governamentais que regulem e monitorem o uso desses compostos.

Nesse sentido, a elaboração de modelos para prever o comportamento de compostos químicos no meio ambiente é uma alternativa muito interessante, pois as informações desejadas podem ser obtidas de forma rápida e barata. Consequentemente, várias pesquisas têm sido realizadas para estudar as relações quantitativas entre a estrutura molecular e as propriedades (QSPR) de tais compostos (BRIGGS, 1981; GAO; GOVIND; TABAK, 1996; LIAO et al., 1996; TAO et al., 1999; BAKER; MIHELCIC; SABLJIC, 2001; KAHN et al., 2005; LIU; YU, 2005; GRAMATICA; GIANI; PAPA, 2007; GOUDARZI et al., 2009; BRONNER; GOSS, 2010; WEN et al., 2012; REIS; SAMPAIO; MELO, 2013; REIS; SAMPAIO; MELO, 2014; SHAO et al., 2014; WANG et al., 2015).

Essa abordagem obteve grande incentivo de organismos internacionais, agências regulatórias governamentais e de entidades públicas e privadas direcionadas à pesquisa e à preservação ambiental (ECHA, 2007, OECD, 2004).

Assim, ressaltou-se neste trabalho, o desenvolvimento de modelos que descrevem o comportamento de substâncias químicas, como um método valioso para avaliação de impactos ambientais e, considerando a diversidade de modelos desenvolvidos, a introdução do uso de um teste estatístico simples para verificar a existência de equivalência entre os modelos (BROWNLEE, 1965).

A estrutura deste texto é composta por esta introdução (seção 1), pela apresentação dos objetivos do trabalho (seção 2), seguida de uma breve revisão bibliográfica sobre o impacto que o uso de produtos químicos causa ao meio ambiente e sobre modelos de relações quantitativas estrutura-propriedade (QSPR) (seção 3) e pela relação de referências citadas nessas seções (seção 4). Destaca-se que o teste adotado, para verificar a equivalência estatística entre modelos de regressão, não foi apresentado na revisão bibliográfica por ser apresentado na seção 2.4 (página 29). Em seguida são apresentados os artigos gerados a partir dos resultados deste trabalho, estando o primeiro em processo de

revisão na revista *Chemosphere*, da editora Elsevier e o segundo em fase final de preparação para submissão. Por fim, a última seção apresenta as considerações finais da pesquisa.

2 OBJETIVOS

2.1 Objetivos gerais

Estudar as relações quantitativas entre a estrutura molecular e as propriedades (QSPR) de agrotóxicos e outros poluentes orgânicos de interesse ambiental para obtenção de modelos que permitam predizer e explicar suas capacidades de contaminação dos solos e águas subterrâneas e, a partir desses modelos, dar subsídios às agências regulatórias para análises preliminares, simples e baratas, sobre a possibilidade do uso de um determinado composto orgânico provocar danos ao meio ambiente.

Propor novos modelos que relacionem o coeficiente de sorção do solo normalizado para o conteúdo de carbono orgânico ($LogK_{oc}$) (variável resposta), a partir de um banco de dados de compostos químicos que contemple uma variedade abrangente de classes de compostos, com descritores moleculares (variáveis explicativas) de fácil e rápida obtenção e que apresentem boa qualidade estatística, boa capacidade de previsão e uma interpretação mecanística adequada à variável resposta sob estudo.

Gerar modelos de $LogK_{oc}$, a partir de valores do logaritmo do coeficiente de partição octanol/água (LogP), obtidos através de algoritmos gratuitos e, adicionalmente, estudar o impacto que o uso de conjuntos de treinamento não tão extensos tem na qualidade de predição dos modelos gerados.

Introduzir o uso de um teste estatístico simples para avaliar se existe equivalência estatística entre os modelos obtidos.

Finalmente, para garantir a confiabilidade dos modelos obtidos, revisar os modelos, considerando os princípios propostos para estudos de QSPR pela Organização para o Desenvolvimento e Cooperação Econômica (OECD, 2004).

2.2 Objetivos específicos

1. Gerar e validar modelos de $LogK_{oc}$, em função do logaritmo do coeficiente de partição octanol/água (LogP), referenciando uma base de dados que considere diversas classes de compostos:
   1.1. Escolher na literatura do banco de dados de coeficientes de sorção no solo de agrotóxicos e outros poluentes orgânicos, obtidos sob mesma metodologia experimental e geração dos valores de LogP associados, mediante o uso de algoritmos gratuitos existentes;
   1.2. Obter modelos QSPR para predição dos coeficientes de sorção:
   1.2.1.1. Usando valores de LogP calculados por algoritmos diferentes;
   1.2.1.2. Usando conjuntos de treinamento de tamanhos diferentes;
   1.3. Analisar a qualidade estatística e validação dos modelos;
   1.4. Avaliar os modelos obtidos, comparativamente a modelos da literatura.
2. Propor o uso de um teste estatístico simples para avaliar a existência de equivalência estatística entre os modelos de regressão visando:
   2.1. Mostrar a equivalência de modelos de $LogK_{oc}$ gerados a partir de valores de LogP calculados por diversos algoritmos gratuítos;
   2.2. Mostrar a equivalência de modelos de $LogK_{oc}$ gerados a partir de conjuntos de treinamento diferentes.

3 REVISÃO BIBLIOGRÁFICA

3.1 Produtos químicos e meio ambiente

A sociedade moderna depende fortemente do uso de produtos químicos para ter uma qualidade de vida alta. Medicamentos, cosméticos, detergentes, pesticidas, tintas, combustíveis, vidros, plásticos, para mencionar alguns exemplos, todos trazem em sua composição produtos químicos. Isso significa que o modelo atual de sociedade seria impensável sem o uso desses compostos. Pensando na grande quantidade de produtos que são descartados diariamente no mundo, pode-se ter uma ideia da quantidade de substâncias químicas que são despejadas. Por essa razão, nas últimas décadas, aumentou a preocupação com o destino final destas substâncias, e os possíveis riscos à saúde humana e ao meio ambiente (MACKAY; WEBSTER, 2003).

Para entender os processos de contaminação, é importante conhecer a natureza química dos grupos funcionais das moléculas que compõem os produtos químico-orgânicos despejados no ambiente. Assim, para compreender os processos de reatividade, toxicidade, degradabilidade e mobilidade desses produtos, nos diversos compartimentos ambientais, é necessário estudar as suas características físicas, químicas e biológicas. As informações sobre esses processos são de suma importância para o planejamento de ações que visem à revitalização ou recuperação de áreas contaminadas (JARDIM; ANDRADE; QUEIROZ, 2009).

O crescimento demográfico, ao mesmo tempo que provoca o aumento das regiões urbanas, restringindo as áreas que podem ser utilizadas para a produção agrícola e pecuária, provoca a necessidade de aumento da produção de alimentos para atender à demanda gerada por esse crescimento. Essa situação levou à proposição de novos métodos e

tecnologias agrícolas que otimizem a produção. Infelizmente, a maior produtividade foi obtida, principalmente, pelo uso crescente de produtos químicos, o que provoca um acúmulo maior de substâncias indesejáveis no solo, provocando o aumento de áreas contaminadas e, consequentemente, dos investimentos que devem ser feitos para a sua recuperação (JARDIM; ANDRADE; QUEIROZ, 2009; PIASAROLO; RIGITANO; GUERREIRO, 2008).

Dentre os compostos químico-orgânicos considerados potencialmente tóxicos aos seres humanos, pode-se destacar os agrotóxicos, substâncias amplamente usadas na agricultura brasileira, com a finalidade de se obter maior produtividade, melhor qualidade das culturas e redução dos custos com mão de obra e energia. Entretanto, apenas 0,1% do que é lançado nas lavouras atinge seus objetivos específicos, o restante tende a mover-se e contaminar os diferentes compartimentos ambientais (ARIAS-ESTÉVEZ et al., 2008).

A alta hidrofobicidade e a baixa reatividade apresentada por esses compostos, associada à ausência de uma forma eficiente de degradação, explicam a tendência de se acumularem em tecidos de organismos vivos (SCHWARZENBACH et al., 1995). Adicionalmente, alguns estudos apontaram a presença dessas substâncias em lençóis freáticos, poços artesianos e minas d’água em áreas agrícolas de diversos países, o que justifica a realização de estudos sobre a mobilidade dessas substâncias no solo (UETA et al., 1999; PIASAROLO; RIGITANO; GUERREIRO, 2008).

No Brasil, a Lei Federal nº 7.802, de 11 de julho de 1989, conhecida como "Lei de Agrotóxicos", adotou e definiu o termo "agrotóxico" para fazer referência às diferentes categorias de uso e, como padronização, sugeriu que, a partir desse momento, fosse adotado pelos autores nos trabalhos publicados na literatura científica nacional (BRASIL, 1989).

Conforme Baird (2002), agrotóxicos podem ser classificados segundo diversos critérios. Considerando o tipo de praga que controlam, pode-se classificá-los como acaricidas, bactericidas, fungicidas, herbicidas, inseticidas, nematicidas, raticidas e vermífugos; Considerando sua estrutura química, pode-se classificá-los como orgânicos (carbamatos, clorados, fosforados e clorofosforados), inorgânicos (compostos que contém arsênio, tálio,

bário, nitrogênio, fósforo, cádmio, ferro, selênio, chumbo, cobre, mercúrio ou zinco) e botânicos: (compostos extraídos de plantas).

No Brasil, como no restante do mundo, as classes de agrotóxicos mais utilizadas são as dos inseticidas, dos herbicidas e dos fungicidas. Dados do Sindicato Nacional da Indústria de Produtos para Defesa Vegetal – SINDIVEG (antigo Sindicato Nacional da Indústria de Produtos para Defesa Agrícola – SINDAG) apontam o Brasil como maior consumidor de agrotóxicos do mundo (LONDRES, 2011). A elevada lucratividade deve ser a grande justificativa para a defesa do uso desses produtos no mercado agrícola nacional. No entanto, o fato do Brasil possuir uma das maiores reservas hídricas do mundo, justifica a necessidade de: 1) pesquisar o destino final dessas substâncias no meio ambiente; 2) propor mecanismos que permitam que agências reguladoras governamentais avaliem o risco de uma substância agredir o meio ambiente.

3.2 Modelos de relações quantitativas estrutura-propriedade (QSPR)

Os estudos das relações quantitativas estrutura-propriedade (*Quantitative Structure-Property Relationship* - QSPR), também chamadas de relações quantitativas estrutura-atividade (*Quantitative Structure-Activity Relationship* - QSAR), visam à construção de modelos que relacionam a estrutura química de uma substância à sua atividade ou a uma propriedade de interesse (DAYAM; NEAMATI, 2003). Tais estudos se baseiam na hipótese de que o comportamento de uma série de substâncias análogas pode ser quantitativamente descrito por modelos matemáticos multiparamétricos. Essa abordagem vem ganhando espaço em vários campos da ciência, principalmente em pesquisas sobre as atividades ou propriedades biológicas, farmacêuticas e ambientais de determinadas classes de compostos químicos, resultando em diversos trabalhos desenvolvidos nas últimas décadas. Exemplos

destes trabalhos são: Dearden (2002), Doucette (2003), Gaudio e Zandonade (2001), Sabljic et al. (1995), Carlsen et al. (2001), Ferreira e Kiralj (2008), Gramatica, Corradi e Consonni (2000), Gramatica e Di Guardo (2002), Todeschini e Consoni (2009), Papa, Dearden e Gramatica (2007), Melagraki et al. (2007), Mitra, Saha e Roy (2011), Kiralj e Ferreira (2009), Reis, Sampaio e Melo (2013), Reis, Sampaio e Melo (2014), Shao et al. (2014), Wang et al. (2015) e Wen et al. (2012).

A construção desses modelos requer a elaboração de um conjunto de dados, geralmente uma matriz (Figura 1), contendo a medida quantitativa da propriedade de interesse (variável resposta, Y) e os parâmetros físico-químicos e estruturais capazes de descrever as propriedades dos compostos (variáveis explicativas, X). Assim, para um grupo de *n* substâncias tem um conjunto de dados contendo os valores da propriedade que será modelada (Y) e as *m* variáveis explicativas X que descrevem essa propriedade.

Y	**X_1**	**X_2**	**...**	**X_m**
Y_1	$X_{1,1}$	$X_{1,2}$	...	$X_{1,m}$
Y_2	$X_{2,1}$	$X_{2,2}$	...	$X_{2,m}$
...	...	...	...	...
Y_n	$X_{n,1}$	$X_{n,2}$	...	$X_{n,m}$

Figura 1 Matriz de dados necessários para a construção de modelos.

Esses modelos devem ser capazes de explicar as relações complexas que existem entre as variáveis independentes (ou explicativas) e as dependentes (ou respostas). Assim, as equações obtidas podem ser utilizadas para predizer as propriedades de outros compostos como, por exemplo, uma atividade toxicológica, farmacêutica ou uma propriedade de interesse ambiental (ERIKSSON et al., 2003; CARBÓ-DORCA; GIRONÉS, 2004).

Em geral, esses modelos são lineares e multidimensionais, e podem ser representados genericamente pela Equação 1, em que $\hat{Y}$ representa os valores previstos da variável resposta; x_1, x_2, … , x_m são as variáveis explicativas (descritores); e b_0, b_1, …, b_m são estimadores dos coeficientes da regressão:

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2 + ... + b_mx_m \quad (1)$$

A qualidade do ajuste do modelo aos valores observados da propriedade de interesse pode ser avaliada pelo cálculo do coeficiente de correlação (r), do desvio-padrão (s) e do teste estatístico F. Em linhas gerais, espera-se que um modelo bem ajustado apresente um valor de r próximo a 1, de s pequeno e de F grande (FERREIRA; MONTANARI; GAUDIO, 2002).

Os descritores de um modelo QSPR refletem propriedades físico-químicas associadas à estrutura molecular dos compostos químicos. Entre as propriedades físico-químicas mais importantes para a descrição de uma atividade ou propriedade de interesse, pode-se citar a distribuição eletrônica, a hidrofobicidade e a estereoquímica da molécula. Cada uma dessas propriedades podem contribuir com maior ou menor intensidade sobre a variável resposta. Portanto, os descritores representam essas propriedades de forma qualitativa e quantitativa, a fim de especificar sua influência na bioatividade de compostos químicos (TAVARES, 2004). Tais variáveis podem ser calculadas utilizando-se diversas abordagens teóricas, como a teoria de grafos, a modelagem molecular ou por meio de metodologias desenvolvidas com base em dados empíricos.

A determinação dos descritores visa obter propriedades moleculares específicas ou padrões de similaridade estrutural entre diferentes moléculas que, por sua vez, podem apresentar uma determinada propriedade ou atividade específica. Alguns tipos de descritores podem permitir, por exemplo, a visualização tridimensional de moléculas, possibilitando a obtenção de informações sobre requisitos estruturais necessários para a interação de uma molécula ativa com seu sítio receptor.

Os descritores podem ser classificados de diversas formas. Ferreira e Kiralj (2008) apresentam um sumário das classificações mais comumente usadas em estudos de QSPR, a saber: i) de acordo com a natureza das unidades estruturais descritas; ii) de acordo com a complexidade em obtê-los, defini-los e descrevê-los; iii) segundo os valores numéricos assumidos; iv) segundo o modo como foram gerados; v) segundo sua dimensionalidade, vi) de acordo com a teoria ou metodologia; vii) de acordo com a sua natureza. Na Tabela 1, é

apresentada a classificação geral formulada por Dunn III (1989), que corresponde à classificação, de acordo com a metodologia proposta por Ferreira e Kiralj (2008).

Tabela 1 Classificação dos descritores que podem ser utilizados em QSPR/ QSAR

Classe	Exemplos
Descritores físico-químicos	Constante eletrônica de Hammet (σ)
	Constante substituinte de campo de Swain Lupton (𝔍)
	Constante substituinte de ressonância Swain Lupton (ℜ)
	Constante hidrofóbica de Hansch (π)
	Constante estérica de Taft (Es)
	Refratividade molar (RM)
	Logaritmo do coeficiente de partição octanol/água (Log P)
	Ponto de fusão
	Momento dipolar (D)
Descritores eletrônicos (mecânico-quânticos)	Cargas atômicas parciais (q)
	Polarizabilidade molecular (α)
	Refratividade molar (RM)
	Energia do orbital molecular HOMO
	Energia do orbital molecular LUMO
	Cargas atômicas parciais
	Parâmetros termodinâmicos
Descritores geométricos	Volume molecular (VM)
	Área molecular (AM)
	Comprimentos e ângulos de ligação
	Ovalidade (Ov)
Descritores topológicos	Índice de Wiener (W)
	Índice de Randic (R)
	Índice de Balanben (B)
	Índices de conectividade (χv)
	Índices de forma Kappa (κ)
Descritores constitucionais	Número de elétrons de valência
	Número de átomos (nAT)
	Número de ligações (nBT)
Combinados	Índices de estado eletrotopológico (E-state indices)
	Densidade de carga superficial

Apesar do conjunto de dados original conter um total de *m* variáveis explicativas (descritores), apenas um subconjunto *k* será utilizado na construção dos modelos. Isso decorre do fato de que existe um limite para o valor de *k* a fim de que haja uma solução única (GELADI; KOWALSKI, 1986). Dessa maneira, o valor máximo de *k* será *m*-1. Porém, a medida que *k* se aproxima de *m* ocorre o sobreajuste ou ajuste forçado (*overfitting*).

Isso significa que o uso de um número excessivo de variáveis explicativas pode dar a falsa impressão de que o modelo está bem ajustado. Em estudos de QSPR, convencionou-se

a inclusão de uma variável explicativa para cada grupo de cinco ou seis compostos do conjunto de dados (FERREIRA; MONTANARI; GAUDIO, 2002).

Como, geralmente, o número total de variáveis disponíveis é muito maior do que o número que será incluído nos modelos, ou seja $m > k$, há a necessidade de usar métodos de seleção das variáveis que são relevantes para a construção dos modelos. Tais métodos ocupam-se em encontrar combinações de k variáveis, dentre as m disponíveis, capazes de produzir modelos que descrevam adequadamente os valores observados da atividade ou propriedade de interesse. Dentre esses métodos, destacam-se a busca sistemática, o algoritmo genético e os métodos baseados na quimiometria (FERREIRA; MONTANARI; GAUDIO, 2002).

As variáveis dependentes, modeladas em QSPR, são atividades ou propriedades químicas, físicas ou biológicas obtidas por meio de estudos experimentais. Muitas vezes, os dados de interesse não possuem distribuição normal, podendo apresentar faixas de variação numérica largas (p.ex., entre 0,1 e 100.000) e concentradas em determinados intervalos. Tal problema pode ser resolvido transformando-se cada valor em seu logaritmo (KUBINYI, 1993).

Dentre os diversos procedimentos matemáticos que podem ser utilizados em estudos QSPR, os mais comuns são a regressão linear múltipla, os métodos de projeção e os métodos não-lineares baseados em redes neurais artificiais.

Na regressão linear múltipla (RLM), os descritores devem ser matematicamente independentes ou ortogonais entre si. De maneira geral, a RLM é utilizada para ajustar um modelo de regressão linear no qual a variável resposta é uma combinação linear das variáveis explicativas (ERIKSSON et al., 2003). Porém, muitas vezes a RLM é inviável devido à existência de multicolinearidade entre os descritores. Nestes casos, os métodos de projeção multivariada são recomendados. Tais métodos são baseados na alteração da dimensionalidade do conjunto de dados.

O conjunto de dados original é transformado, de modo a se obter um novo conjunto com um número bem menor de variáveis independentes. As novas variáveis são obtidas pelas

combinações lineares das variáveis originais. Uma das vantagens é que as novas variáveis são ortogonais entre si, o que soluciona o problema da multicolinearidade da RLM (LIVINGSTONE, 2003).

Dois métodos de regressão utilizam esse princípio: a regressão por componentes principais (*Principal Components Regression* -PCR) e a regressão por quadrados mínimos parciais (*Partial Least Squares* - PLS). No primeiro, as novas variáveis são chamadas componentes principais (*Principal Components* - PC) e são formadas por combinações lineares das variáveis originais. No segundo, as variáveis são denominadas variáveis latentes (*Latent Variables* - LV). Enquanto no método PCR, as PCs são construídas primeiro e apenas depois é realizada a regressão perante a variável resposta Y, no método PLS as LVs são obtidas utilizando-se dos valores de Y no processo de sua obtenção. Assim, espera-se que a covariância entre as coordenadas das amostras no novo sistema de eixos e Y seja maximizada, com a obtenção de resultados superiores àqueles que seriam obtidos com PCR (FERREIRA; KIRALJ, 2008; ROY; ROY, 2008).

Além dos métodos de regressão linear RLM, PCR e PLS, outros podem ser utilizados quando se deseja estudar as relações quantitativas não-lineares. Um método muito usado para tal finalidade é o das redes neurais artificiais (*Artificial Neural Networks*- ANN). Essa metodologia utiliza pesquisas de inteligência artificial que buscam replicar a estrutura do cérebro, simulando suas funções. Assim, uma ANN consiste em camadas de neurônios artificiais conectados uns aos outros (LIVINGSTONE, 2003).

Porém, ao contrário dos métodos lineares tradicionais, na ANN não existe como se determinar a rede neural ótima para o ajuste de um conjunto de dados. Assim, normalmente os algoritmos são utilizados durante certo número de vezes até a seleção da melhor rede (AGATONOVIC-KUSTRIN; TURNER; GLASS, 2008).

Adicionalmente, é importante destacar que todo modelo QSPR deve ser validado antes de sua interpretação e utilização na predição das propriedades de interesse (WOLD; ERIKSSON, 1998). Segundo o guia da Organização para a Cooperação e Desenvolvimento

Econômico (*Organization for Economic Cooperation and Development* – OECD), a validação é definida como “o processo pelo qual a confiabilidade e a relevância de uma abordagem, método, processo ou avaliação específica, são estabelecidas com um objetivo definido” (OECD, 2004. Portanto, a aplicabilidade e confiabilidade de modelos QSPR estão relacionadas com a sua aprovação em testes de validação específicos da área.

Para finalizar, a Figura 2 apresenta o procedimento proposto por Gramatica (2013) para o desenvolvimento de modelos QSPR com qualidade estatística alta e preditivos. Este procedimento é o que foi, parcialmente, utilizado neste estudo e visa sintetizar os passos necessários para o desenvolvimento e validação de modelos QSPR.

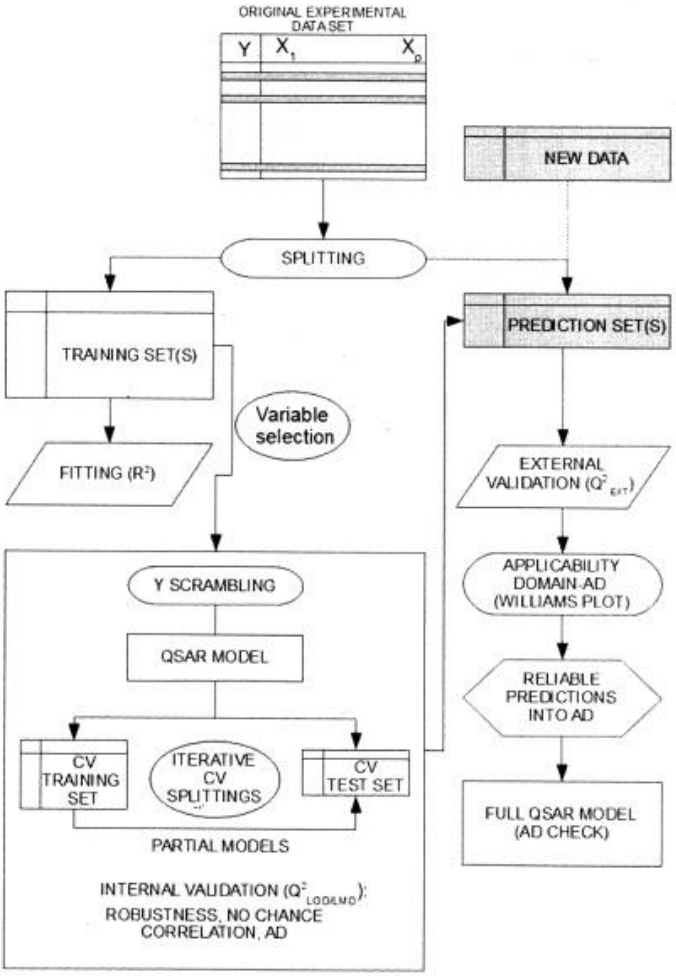


Figura 2 Procedimento para o desenvolvimento de modelos QSPR preditivos.

Fonte: Adaptado de Gramática (2013).

REFERÊNCIAS

AGATONOVIC-KUSTRIN, S.; TURNER, J. V.; GLASS, B. D. Molecular structural characteristics as determinants of estrogen receptor selectivity. **Journal of Pharmaceutical and Biomedical Analysis**, v. 48, n. 2, p. 369–375, 2008.

ARIAS-ESTÉVEZ, M.; LOPEZ-PERIAGO, E.; MARTINEZ-CARBALLO, E.; SIMAL-GANDARA, J.; MEJUTO, J.C.; GARCIA-RIO, L. The mobility and degradation of pesticides in soils and the pollution of groundwater resources. **Agriculture, Ecosystems & Environment**, v. 123, n. 4, p. 247-260, 2008.

BAIRD, C. **Química ambiental.** 2. ed. Porto Alegre: Bookman, 2002.

BAKER, J. R.; MIHELCIC, J. R.; SABLJIC, A. Reliable QSAR for estimating Koc for persistent organic pollutants: correlation with molecular connectivity indices. **Chemosphere**, v. 45, p. 213-221, 2001.

BOLOGNESI, C. Genotoxicity of pesticides: a review of human biomonitoring studies. **Mutation Research/Reviews in Mutation Research**, v. 543, n. 3, p. 251-272, 2003.

BRASIL. **Lei Federal nº 7.802**, de 11 de julho de 1989. Dispõe sobre a pesquisa, a experimentação, a produção, a embalagem e rotulagem, o transporte, o armazenamento, a comercialização, a propaganda comercial, a utilização, a importação, a exportação, o destino final dos resíduos e embalagens, o registro, a classificação, o controle, a inspeção e a fiscalização de agrotóxicos, seus componentes e afins, e dá outras providências. Brasília. 1989. Disponível em: http://www.planalto.gov.br/CCIVIL_03/LEIS/L7802.htm. Acesso em: 21 out. 2009.

BRIGGS, G. G. Theoretical and experimental relationships between soil adsorption, octanol-water partition coefficients, water solubilities, bioconcentration factors, and the parachor. **Journal of Agricultural and Food Chemistry**, v**.** 29, n. 5, p. 1050-1059, 1981.

BRONNER, G.; GOSS, K. U. Predicting sorption of pesticides and other multifunctional organic chemicals to soil organic carbon. **Environmental Science & Technology**, v. 45, p. 1313-1319, 2010.

BROWNLEE, K. A. **Statistical theory and methodology in science and engineering**. New York: John Wiley & Sons, 1965.

CARBÓ-DORCA, R.; GIRONÉS, X. Quantum similarity and quantitative structure-activity relationships. *In:* BULTINCK, P.; WINTER, H.; LANGENAEKER, W.; TOLLENAERE, J. P.; (Orgs). **Computational medicinal chemistry for drug design.** New York: Marcel Dekker Inc., 2004.

CARLSEN, L.; SORENSEN, P. B.; THOMSEN, M. Partial order ranking-based QSAR´s: estimation of solubilities and octanol-water partitioning. **Chemosphere**, v. 43, p. 295-302, 2001.

DAYAM, R.; NEAMATI, N. Small-molecule HIV-1 integrase inhibitors: the 2001-2002 update. **Current Pharmaceutical Design**, v. 9, n. 2, p. 1789-1802, 2003.

DEARDEN, J. C. Prediction of enviromental toxicity and fate using quantitative structure-activity relationships (QSARs). **Journal of Brazilian Chemistry Society**, v. 3, n. 6, p. 754-762, 2002.

DOUCETTE, W. J. Quantitative structure-activity relationships for predicting soil/sediment sorption coefficients for organic chemicals. **Environmental Toxicology and Chemistry**, v. 22, n. 8, p. 1771-1788, 2003.

DUNN III, W. J. Quantitative structure-activity relationships (QSAR). **Chemometrics and Intelligent Laboratory Systems**, v. 6, n. 3, p. 181-190,1989.

ERIKSSON, L.; JAWORSKA, J.; WORTH, A. P.; CRONIN, M. T. D.; MCDOWELL, R. M.; GRAMÁTICA, P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. **Environmental Health Perspectives**, v. 111, n. 10, p. 1361-1374, 2003.

EUROPEAN CHEMICALS AGENCY - ECHA. **Registration, evaluation, authorisation and restriction of chemicals** - REACH. 2007. Disponível em: http://ec. europa. eu/environment/chemicals/reach/reach_intro. html. Acesso em: 27/04/2014.

FERREIRA, M. M. C.; MONTANARI, C. A.; GAUDIO, A. C. Seleção de variáveis em QSAR. **Química Nova,** São Paulo - SP, v. 25, n. 3, p. 439-448, 2002.

FERREIRA, M. M. C.; KIRALJ, R. Métodos quimiométricos em relações quantitativas estrutura-atividade (QSAR). *In*: MONTANARI, Carlos A. (Org.). **Química medicinal**: métodos e fundamentos em planejamento de fármacos. 1. ed. Viçosa: Editora UFV, 2008.

GAO, C.; GOVIND, R.; TABAK, H. H. Predicting soil sorption coefficients of organic chemicals using a neural network model. **Environmental Toxicology and Chemistry**, v. 15, p. 1089-1096, 1996.

GAUDIO, A. C.; ZANDONADE, E. Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica. **Química Nova**, São Paulo - SP, v. 24, n. 5, p. 658-671, 2001.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, n. C, p. 1-17, 1986.

GOUDARZI, N.; GOODARZI, M.; ARAUJO, M. C. U.; GALVÃO, R. K. H. QSPR modeling of soil sorption coefficients (Koc) of pesticides using SPA-ANN and SPA-MLR. **Journal of Agricultural and Food Chemistry**, v. 57, p. 7153-7158, 2009.

GRAMATICA, P.; CORRADI, M.; CONSONNI, V. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. **Chemosphere**, v. 41, n. 5, p. 763-777, 2000.

GRAMATICA P.; DI GUARDO, A. Screening of pesticides for environmental partitioning tendency. **Chemosphere**, v. 47, p. 947–956, 2002.

GRAMATICA, P.; GIANI, E.; PAPA, E. Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. **Journal of Molecular Graphics and Modelling**, v. 25, p. 755-766, 2007.

GRAMATICA, P. On the development. and validation of QSAR models. *In*: REISFELD, B.; MAYENO, A. N. (Eds.).**Computational Toxicology**: New York: Human Press, 2013. Volume II. (Series: Methods in molecular biology, v. 930).

JARDIM, I. C. S. F.; ANDRADE, J. A.; QUEIROZ, S. C. N. Resíduos de agrotóxicos em alimentos: uma preocupação ambiental global - Um enfoque às maçãs. **Química Nova**, São Paulo – SP, v. 32, n. 4, p. 998-1012, 2009.

JURASKE, R.; ANTÓN, A.; CASTELLS, F.; HUIJBREGTS, M. A. J. Pestscreen: a screening approach for scoring and ranking pesticides by their environmental and toxicological concern. **Environment International**, v. 33, p. 886–893, 2007.

KAHN, I.; FARA, D.; KARELSON, M.; MARAN, U.; ANDERSSON, P. L. QSPR treatment of the soil sorption coefficients of organic pollutants. **Journal of Chemical Information and Modeling**, v. 45, p. 94-105, 2005.

KIRALJ, R.; FERREIRA, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. **Journal of the Brazilian Chemical Society**, v. 20, n. 4, p. 770-787, 2009.

KUBINYI, H. **QSAR**: hansch analysis and related approaches.Weinheim, Germany: VCH, 1993.

LIAO, Y. Y.; WANG, Z. T.; CHEN, J. W.; HAN, S. K.; WANG, L. S.; LU, G. Y.; ZHAO, T. N. The prediction of soil sorption coefficients of heterocyclic nitrogen compounds by octanol/water partition coefficient, water solubility, and by molecular connectivity indices. **Bulletin of Environmental Contamination and Toxicology**, v. 56, p. 711-716, 1996.

LIVINGSTONE, D. J. Quantitative structure-activity relationships. *In*: KING, F. D. (Org.). **Medicinal Chemistry: Principles and Practice**. 2. ed. Cambridge: Royal Society of Chemistry, 2003.

LIU, G.; YU, J. QSAR analysis of soil sorption coefficients for polar organic chemicals: substituted anilines and phenols. **Water Research**, v. 39, p. 2048-2055, 2005.

LONDRES, F. **Agrotóxicos no Brasil**: um guia para ação em defesa da vida. AS-PTA : Rio de Janeiro. 2011. Disponível em: aspta.org.br/wp-content/uploads/2011/09/Agrotoxicos-no-Brasil-mobile.pdf. Acesso em: 21 out. 2016.

MACKAY, D.; WEBSTER, E. A Perspective on environmental models and QSARs. **SAR and QSAR in Environmental Research**, v. 14, n. 1, p. 7-16, 2003.

MELAGRAKI, G.; AFANTITIS, A.; SARIMVEIS, H.; KOUTENTIS, P. A.; MARKOPOOLUS, J.; IGGLESSI-MARKOPOULOU, O. Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening. **Journal of Computer-Aided Molecular Design**, v. 21, n. 1, p. 251-267, 2007.

MITRA, I.; SAHA, A.; ROY, K. Chemometric QSAR modeling and in silico design of antioxidant NO donor phenols. **Scientia Pharmaceutica**, v. 79, n. 1, p. 31-57, 2011.

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT - OECD. **Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models**. Paris: OECD, 2004.

PAPA, E.; DEARDEN, J. C.; GRAMATICA, P. Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. **Chemosphere**, v. 67, p. 351–358, 2007.

PIASAROLO, L.; RIGITANO, R. L. O.; GUERREIRO, M. C. Influência da polaridade de pesticidas não-iônicos sobre sua sorção em um latossolo. **Ciência e Agrotecnologia**, Lavras - MG, v. 32, n. 6, p. 1802-1809, 2008.

REIS, R. R.; SAMPAIO, S. C.; MELO, E. B. An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficient. **Water Research**, v. 53, p. 191-199, 2014.

REIS, R. R.; SAMPAIO, S. C.; MELO, E. B. The effect of different log P algorithms on the modeling of the soil sorption coefficient of nonionic pesticides. **Water Research**, v. 47, p. 5751-5759, 2013.

ROY, P.; ROY, K. On some aspects of variable selection for partial least squares regression models. **QSAR & Combinatorial Science**, v. 27, n. 3, p. 302-313, 2008.

SABLJIC, A.; GUSTEN, H.; VERHAAR, H.; HERMENS, J. QSAR modelling of soil sorption: Improvements and systematics of log K_{OC} vs log K_{OW} correlations. **Chemosphere**, v. 31, n. 11, p. 4489-4514, 1995.

SCHWARZENBACH, R. P.; GSCHWEND, P. M.; IMBODEN, D. M. **Environmental organic chemistry.** Wiley-Interscience: USA, 1995.

SHAO, Y.; LIU, J.; WANG, M.; SHI, L.; YAO, X.; GRAMATICA, P. Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods. **Atmospheric Environment**, v. 88, p. 212–218, 2014.

TAO, S.; PIAO, H.; DAWSON, R.; LU, X.; HU, H. Estimation of organic carbon normalized sorption coefficient (Koc) for soils using the fragment constant method. **Environmental Science & Technology**, v. 33, p. 2719-2725, 1999.

TAVARES, L. C. QSAR: a abordagem de Hansch. **Química Nova**, São Paulo - SP, v. 27, n. 4, p. 631-639, 2004.

TODESCHINI, R.; CONSONI, V. **Molecular descriptors for chemoinformatics**. Weinheim, Germany: Wiley-VCH, 2009.

UETA, J.; CERDEIRA, A. L.; PEREIRA, N. L.; SHUHAMA, I. K. Biodegradação de herbicidas e biorremediação. **Biotecnologia Ciência e Desenvolvimento**, Brasília - DF, v. 2, n. 10, p. 10-13, 1999.

WANG, Y.; CHEN, J.; YANG, X.; LYAKURWA, F.; LI, X.; QIAO, X. *In silico* model for predicting soil organic carbon normalized sorption coefficient (Koc) of organic chemicals. **Chemosphere**, v. 119, p. 438-444, 2015.

WEN, Y.; SU, L. M.; QIN, W. C.; FU, L.; HE, J.; ZHAO, Y. H. Linear and non-linear relationships between soil sorption and hydrophobicity: model, validation and influencing factors. **Chemosphere**, v. 86, p. 634–640, 2012.

WOLD, S.; ERIKSSON, L. Statistical validation of QSAR results. *In*: VAN DER WATERBEEMD, H. (Org.). **Chemometric methods in molecular design.** Weinheim, Germany: VCH, 1998.

4 ARTIGOS

ARTIGO 1 – EQUIVALÊNCIA ESTATÍSTICA DE MODELOS DE PREDIÇÃO DO COEFICIENTE DE SORÇÃO DO SOLO OBTIDOS A PARTIR DE DIFERENTES ALGORITMOS DE LOGP[1]

1 INTRODUÇÃO

Atualmente, há uma grande quantidade de substâncias químicas, produzidas para fins diversos, que são lançadas no meio ambiente. Tais produtos podem causar danos ambientais e também à saúde humana. Por essas razões, o monitoramento desses produtos é necessário, para evitar riscos potenciais. (MACKAY et al., 2006).

O desenvolvimento de modelos matemáticos para predição de propriedades físicas, químicas e biológicas desses compostos reduz o custo e o tempo gasto na determinação experimental dos parâmetros de interesse ambiental. Dessa maneira, estudos que mensuram as relações quantitativas entre a estrutura molecular e as propriedades das substâncias (QSPR) permitem explicar e avaliar a capacidade de contaminação dos solos e das águas pelo uso de agrotóxicos e outros poluentes orgânicos (HANSCH; LEO; HOEKMAN, 1995).

Considerando que o coeficiente de sorção do solo normalizado para o conteúdo de carbono orgânico (Koc) é um importante parâmetro físico-químico que pode ser utilizado para determinar o destino final das substâncias químicas lançadas na natureza (DOUCETTE, 2003;

[1] Artigo submetido a revista *Chemosphere* (Oxford) – ISSN: 0045-6535. Classificação A1, em Periódicos Qualis 2015 - Ciências Agrárias I.

HUUSKONEN, 2003), um grande número de modelos para prever $LogK_{oc}$ tem sido desenvolvido com a intenção de se obter uma base de dados atualizada e confiável (SABLJIC et al., 1995; GRAMATICA; CORRADI; CONSONNI, 2000; TAO et al., 2001; SCHÜÜRMANN; EBERT; KÜHNE, 2006; GRAMATICA; GIANI; PAPA, 2007; GOUDARZI et al., 2009; WEN et al., 2012; REIS; SAMPAIO; MELO, 2014; SHAO et al., 2014; WANG et al., 2015). Além desses estudos, diferentes abordagens matemáticas têm sido usadas na obtenção de modelos, tais como: *Linear Regression* (LR), *Multiple Linear Regression* (MLR), *Local Lazy Regression* (LLR), *Principal Component Regression* (PCR) e *Partial Least Squares* (PLS) (ROWE, 2010; GRAMATICA, 2013).

Diversos desses modelos foram desenvolvidos com base na relação entre $LogK_{oc}$ e o logaritmo do coeficiente de partição octanol/água (LogP), sendo este último amplamente utilizado para descrever o comportamento hidrofílico ou hidrofóbico de um composto (GAWLIK et al., 1997; DOUCETTE, 2003; CRONIN; LIVINGSTONE, 2004; RAZZAQUE; GRATHWOHL, 2008; REIS; SAMPAIO; MELO, 2013).

Assim, o parâmetro LogP é uma variável importante a ser considerada na modelagem de $LogK_{oc}$, uma vez que a hidrofobicidade é a força motriz do processo de sorção no solo das substâncias não iônicas, pois as interações das moléculas destes compostos ocorrem majoritariamente com a matéria orgânica do solo (ALLEN-KING; GRATHWOHL; BALL, 2002; DOUCETTE, 2003; WEN et al., 2012).

Embora o ideal fosse a utilização de valores experimentais de LogP para se obter modelos QSPR mais realistas, os dados experimentais de muitos compostos não são disponíveis. Assim, abordagens computacionais para calcular LogP são ferramentas muito usadas em estudos QSPR. Atualmente, existem vários algoritmos, comerciais e gratuitos, disponíveis para calcular LogP (MANNHOLD et al., 2009; REIS et. al., 2013), cujas metodologias usadas nos cálculos podem ser encontradas na literatura científica (MANNHOLD; VAN DE WATERBEEMD, 2001; TETKO; TANCHUK; VILLA, 2001; CHENG et al., 2007; TODESCHINI; CONSONNI, 2009).

O estudo de Reis et al. (2013) chama a atenção para o fato de que estudos QSPR sobre o coeficiente de sorção do solo não detalham como ou porque um determinado algoritmo para cálculo de LogP foi escolhido ou se algum critério objetivo para a escolha desse algoritmo assegura que o melhor modelo de regressão será gerado. Assim, Reis et al. (2013) avaliaram diferentes algoritmos gratuitos para cálculo de LogP na modelagem de $LogK_{oc}$, para determinar qual desses algoritmos seria mais adequado para esses fins, apresentando como resultado um ranking. No entanto, esse trabalho limitou o conjunto de compostos utilizados a pesticidas não iônicos. Por isso, estudos mais abrangentes, envolvendo outras classes de compostos, seriam desejáveis. Por outro lado, uma questão importante a ser respondida é se existe equivalência estatística entre modelos obtidos a partir de diferentes algoritmos de LogP, pois a existência dessa equivalência permitiria que, na impossibilidade de se obterem dados de LogP gerados por um algoritmo, sejam utilizados dados gerados por outro algoritmo.

Nesse sentido, o objetivo deste estudo foi: i) utilizando um conjunto de dados amplo e diversificado (SHAO et al., 2014), avaliar os algoritmos gratuitos para calcular LogP na modelagem de $LogK_{oc}$, para determinar quais desses algoritmos são os mais adequados; ii) introduzir um teste estatístico simples para avaliar se há equivalência estatística entre os melhores modelos obtidos. Ressalta-se que todos os modelos obtidos foram testados e validados de acordo com as diretrizes da literatura (KIRALJ; FERREIRA, 2009; CHIRICO; GRAMATICA, 2011; ROY et al., 2012), para garantir que são confiáveis e bons para predição.

2 MATERIAIS E MÉTODOS

2.1 Valores experimentais de $LogK_{oc}$

Os valores experimentais do coeficiente de sorção do solo ($LogK_{oc}$) utilizados neste trabalho foram extraídos de Shao et al. (2014). O conjunto foi selecionado por ser extenso (contém 964 compostos orgânicos não iônicos) e heterogêneo, pois apresenta compostos de diferentes classes químicas. Os valores experimentais do $LogK_{oc}$ usados são apresentados no material suplementar (Tabela 1 - Apêndice A).

Os dados foram divididos em dois conjuntos: i) um conjunto de treinamento, formado por 643 compostos, para construir os modelos QSPR; ii) um conjunto de validação externa, formado por 321 compostos, para avaliar a capacidade de predição dos modelos obtidos. Esses conjuntos foram os mesmos utilizados por Shao et al. (2014). Assim, será possível fazer comparações com os resultados obtidos nesta pesquisa.

2.2 Obtenção dos valores de LogP e dos modelos QSPR

Os valores do coeficiente de partição octanol/água (LogP) de todos os compostos foram obtidos utilizando-se diferentes algoritmos propostos na literatura, a saber: ALOGPs, AC_logP, ALOGP, MLOGP, KOWWIN, XLOGP2 e XLOGP3. Detalhes sobre as características destes algoritmos podem ser consultados em Mannhold e van de Waterbeemd (2001), Tetko et al. (2001); Cheng et al. (2007) e Todeschini e Consonni (2009).

Os valores de LogP foram calculados para cada composto a partir do seu número *Chemical Abstracts Service* (CAS) ou do seu *Simplified Molecular Input Line Entry System* (SMILES), utilizando-se o programa ALOGPS 2.1 do *Virtual Computational Chemistry Laboratory*, disponibilizado em http://www.vcclab.org/lab/alogps/.

O *software Estimation Programs Interface Suite*TM (EPI SuiteTM), desenvolvido pelo Escritório de Prevenção da Poluição e Tóxicos da *US Environmental Protection Agency* (EPA) e pela *Syracuse Research Corporation* (SRC), foi utilizado para obter os SMILES dos compostos e, também, os valores de KOWWIN de alguns compostos que não foram disponibilizados pelo programa ALOGPS 2.1. Assim, ao final do processo, obteve-se uma tabela com o valor de $LogK_{oc}$ e todos os valores de LogP associados, para cada composto (Tabela 1 – Apêndice A).

Os modelos para estimar o $LogK_{oc}$ em função do LogP foram obtidos por regressão linear simples, utilizando-se o Minitab 17.2.1 (Minitab Inc., USA) e QSARINS 2.2.1 (GRAMATICA et al., 2013; GRAMATICA et al., 2014). Inicialmente foram feitas regressões com todos os compostos do conjunto de treinamento (n = 643). Em seguida, 4 compostos (glyphosate, p-benzidine, ciprofloxacin, enrofloxacin) foram excluídos do conjunto, pois eram pontos discrepantes para todos os modelos e apresentavam resíduo padronizado (SR) maior que 5,50. Novas regressões foram feitas, sendo gerados sete modelos, cada um utilizando um dos algoritmos de LogP. A qualidade estatística foi avaliada e os modelos foram comparados entre si.

2.3 Qualidade estatística e poder de predição dos modelos

A qualidade do ajuste de um modelo é auferida a partir da avaliação de quão bem a especificação do modelo se ajusta aos dados experimentais. Essa avaliação é realizada pelo

cálculo do coeficiente de determinação (R^2), da soma dos quadrados do erro residual previsto do conjunto de treinamento (RSS_{tr}), da raiz do erro quadrado médio do conjunto de treinamento ($RMSE_{tr}$) e do coeficiente de concordância da correlação para o conjunto de treinamento (CCC_{tr}). Para que um modelo seja considerado bem ajustado, precisa ter valores de R^2 maiores que 0,7 e os valores de RSS_{tr} e de $RMSE_{tr}$ devem ser os menores possíveis; o valor de CCC_{tr} deve ser maior do 0,85 (CHIRICO; GRAMATICA, 2011; 2012; ROY et al., 2012; GAUDIO; ZANDONADE, 2001).

A significância dos modelos foi avaliada através da estatística F. O valor da estatística F da regressão deve ser maior do que um valor de referência tabelado a 5% de nível de significância (α=0,05) (KIRALJ; FERREIRA, 2009).

A confiabilidade estatística dos modelos foi avaliada através de validação interna ou validação cruzada, *Leave-One-Out* (LOO). Para realizar essa validação, exclui-se um a um, cada objeto do modelo; reconstrói-se o modelo sem o objeto excluído e calcula-se o valor desse objeto. Finalmente, calculam-se a soma dos quadrados das diferenças entre os valores preditos e observados ($PRESS_{cv}$), o coeficiente de determinação da validação cruzada (Q^2_{LOO}) e a raiz do erro quadrado médio da validação cruzada ($RMSE_{cv}$). Segundo Chirico e Gramatica (2011; 2012), um modelo pode ser considerado estatisticamente confiável se o valor de $RMSE_{cv}$ for próximo de 0 e o valor de Q^2_{LOO} for maior que 0,6. Além desses parâmetros, foi calculado o coeficiente de concordância da correlação da validação cruzada (CCC_{cv}), que deve ser maior do que 0,85.

Para avaliar a estabilidade dos modelos frente a pequenas mudanças nos seus parâmetros, utilizou-se a validação cruzada *Leave-Many-Out* (LMO). Essa técnica iterativa exclui, aleatoriamente, uma porcentagem de compostos do conjunto de treinamento. Neste estudo foram realizadas 2000 iterações e, a cada iteração, excluídos 30% dos compostos. Segundo Kiralj e Ferreira (2009), para que o modelo seja considerado robusto, o valor médio de Q^2_{LMO} deve estar o mais próximo possível do valor de Q^2_{LOO}.

A possibilidade de correlação ao acaso pode ser testada por Y-randomização (RÜCKER; RÜCKER; MERINGER, 2007). Esse teste avalia se a relação entre as variáveis explicativas e a variável resposta foi resultado do acaso. Essa técnica calcula iterativamente um número determinado de modelos embaralhando, aleatoriamente, os valores da variável resposta. Para descartar a possibilidade de correlação ao acaso do modelo avaliado, os valores de R^2 e de Q^2_{LOO} devem ser maiores do que R^2_{Yscr} e Q^2_{Yscr}, respectivamente, e o valor de $RMSE_{cv}$ menor do que $RMSE\_AV_{Yscr}$.

A avaliação do poder de predição dos modelos de regressão é realizada pela validação externa, que foi feita mediante a predição dos valores de $LogK_{oc}$ para o conjunto de teste (321 compostos). Assim, o poder de predição dos modelos pode ser avaliado a partir do coeficiente de determinação da validação externa (R^2_{ext}) e do coeficiente de determinação da validação externa modificado (r^2_m) (Ojha et al., 2011). Segundo Chirico e Gramatica (2011, 2012), os valores de R^2_{ext} devem ser maiores do que 0,7, o valor de $\overline{r^2_m}$ deve ser maior do que 0,65 e o valor de Δr^2_m deve ser menor do que 0,2. Foram determinados ainda os valores do desvio padrão da predição ($RMSE_{ext}$) e da soma dos quadrados dos resíduos das predições da validação externa ($PRESS_{ext}$), os quais devem ser os menores possíveis. O valor do coeficiente de concordância da correlação (CCC_{ext}) deve ser maior do que 0,85 para que o modelo seja adotado (LIN, 1989; CHIRICO; GRAMATICA, 2011; 2012).

Finalmente, após esses procedimentos de validação, os melhores modelos obtidos foram comparados entre si, para verificar existência de equivalência estatística entre os modelos. As fórmulas usadas para calcular os parâmetros estatísticos mencionados são apresentadas no material suplementar (Tabelas 2, 3 e 4 – Apêndice A).

2.4 Teste de equivalência estatística entre os modelos

Neste estudo, foi proposta a utilização do procedimento descrito por Brownlee (1965), para verificar se os modelos são estatisticamente equivalentes. O teste consiste em verificar, para modelos obtidos a partir de diferentes conjuntos de dados, se as variâncias são iguais, se há paralelismo entre as retas de regressão e se os interceptos dos modelos são iguais. Caso essas três igualdades sejam constatadas, os modelos são considerados equivalentes.

Assim, considerando dois grupos de observação (x11, y11), (x12, y12), ..., (x1n1, y1n1), n1 pares de dados, e (x21, y21), (x22, y22), ..., (x2n2, y2n2), n2 pares de dados, é possível calcular os parâmetros dos modelos ajustados pelo método dos mínimos quadrados, de modo a obter as equações: $\hat{y}_{1} = \hat{\beta}_{01} + \hat{\beta}_{11}X$ e $\hat{y}_{2} = \hat{\beta}_{02} + \hat{\beta}_{12}X$.

Para verificar a igualdade das variâncias, as hipóteses $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ vs $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$ são testadas a partir da seguinte estatística:

$$F_1 = \frac{Maior\{S_1^2, S_2^2\}}{Menor\{S_1^2, S_2^2\}} \sim F(n_1 - 2; n_2 - 2) \quad (1)$$

em que: $S_1^2 = MSE$ da reta ajustada Y1 e $S_2^2 = MSE$ da reta ajustada Y2 e $\mathrm{F}(\mathrm{n}_1\text{-}2; \mathrm{n}_2\text{-}2) = \mathrm{F_c}$ é o ponto crítico da tabela F-Snedecor, a 5% de significância, com $\mathrm{n}_1\text{-}2; \mathrm{n}_2\text{-}2$ graus de liberdade no numerador e denominador respectivamente. A hipótese H_0 é aceita se $F_1 < F_c$.

Para verificar o paralelismo entre as retas de regressão, comparou-se os coeficientes angulares, através das hipóteses $H_0 : \hat{\beta}_{11} = \hat{\beta}_{12}$ vs $H_1 : \hat{\beta}_{11} \neq \hat{\beta}_{12}$.

Como $\hat{\beta}_{11}$ e $\hat{\beta}_{12}$ têm distribuição normal e são variáveis independentes, a variância estimada da diferença de $\hat{\beta}_{11} - \hat{\beta}_{12}$ tem a forma:

$$Var(\hat{\beta}_{11} - \hat{\beta}_{12}) = S^2 \left[\frac{1}{(n_1 - 1)S_{x1}^2} + \frac{1}{(n_2 - 1)S_{x2}^2}\right] \quad (2)$$

em que: $S^2 = \frac{(n_1-2)S_1^2+(n_2-2)S_2^2}{n_1+n_2-4}$, S_{x1}^2 é a variância de X para a população 1 e S_{x2}^2 é a variância de X para a população 2.

Assim, para testar a hipóteseH_0, usa-se a estatística:

$$T_1 = \frac{\hat{\beta}_{11}-\hat{\beta}_{12}}{[Var(\hat{\beta}_{11}-\hat{\beta}_{12})]^{1/2}} \sim t(n_1+n_2-4) \quad (3)$$

em que: $t(n_1+n_2-4) = t_c$ é o valor crítico da tabela t-Student bicaudal com n_1+n_2-4 graus de liberdade e nível de 5% de significância. Assim, o paralelismo entre as retas é verificado quando $|T_1| < t_c$.

Finalmente, é necessário verificar se os interceptos são iguais. Isto permite comprovar se as retas são coincidentes. A hipótese apropriada para verificar a igualdade dos interceptos é $H_0 : \hat{\beta}_{01} = \hat{\beta}_{02}\ vs\ H_1 : \hat{\beta}_{01} \neq \hat{\beta}_{02}$. A estatística do teste, sob H_0, é:

$$T_2 = \frac{\hat{\beta}_{01}-\hat{\beta}_{02}}{[Var(\hat{\beta}_{01}-\hat{\beta}_{02})]^{1/2}} \sim t(n_1+n_2-3) \quad (4)$$

em que: $Var(\hat{\beta}_{01}-\hat{\beta}_{02}) = S^2\left[\frac{1}{n_1}+\frac{1}{n_2}+\frac{\bar{X}_1^2}{(n_1-1)S_{x1}^2}+\frac{\bar{X}_2^2}{(n_2-1)S_{x2}^2}\right]$ e $t(n_1+n_2-3) = t_c$ é o valor crítico da tabela t-Student bicaudal, com n_1+n_2-3 graus de liberdade, a 5% de significância. Rejeita-se H_0, a 5% de significância, se $|T_2| \geq t_c$. Caso contrário, se $|T_2| < t_c$, aceita-se a hipótese nula e pode-se concluir que os interceptos são iguais ao nível de 5% de significância.

Assim, aceitando as três hipóteses nulas ($H_0 : \sigma_1^2 = \sigma_2^2$; $H_0 : \hat{\beta}_{11} = \hat{\beta}_{12}\ e\ H_0 : \hat{\beta}_{01} = \hat{\beta}_{02}$), verifica-se que os modelos são estatisticamente equivalentes.

3 RESULTADOS E DISCUSSÃO

3.1 Modelos QSPR de predição de $LogK_{oc}$

A Tabela 1 apresenta os modelos de estimação do $LogK_{oc}$ calculados pelo QSARINS, considerando os algoritmos para cálculo do LogP utilizados neste estudo. A Tabela 2 mostra os parâmetros necessários para avaliar a qualidade estatística dos modelos.

Tabela 1 Modelos de predição de $LogK_{oc}$ (n = 639)

Modelo	Equação
M1	$LogK_{oc}$ = 1,322 + 0,530 ALOGPs
M2	$LogK_{oc}$ = 1,216 + 0,572 AC_logP
M3	$LogK_{oc}$ = 1,284 + 0,585 ALOGP
M4	$LogK_{oc}$ = 1,281 + 0,600 MLOGP
M5	$LogK_{oc}$ = 1,308 + 0,534 KOWWIN
M6	$LogK_{oc}$ = 1,326 + 0,542 XLOGP2
M7	$LogK_{oc}$ = 1,293 + 0,545 XLOGP3

Todos os modelos apresentaram-se estatisticamente significativos, a 5% de significância, pois os valores da estatística F são maiores do que o valor de F tabulado ($F_{1,637}$=3,86). Os modelos M1, M5 e M7, que utilizam os algoritmos ALOGPs, KOWWIN e XLOGP3, respectivamente, mostraram-se mais significativos, pois apresentaram os maiores valores (Tabela 2). Em relação aos valores do coeficiente de determinação (R^2), verificou-se que os mesmos três modelos apresentam os melhores ajustes.

Adicionalmente, observou-se que os três modelos são os que possuem os menores valores de $RMSE_{tr}$ e RSS_{tr} e os maiores valores de CCC_{tr}, o que mostra que são os modelos que possuem menor erro e menor diferença entre os dados experimentais e os preditos.

Tabela 2 Parâmetros estatísticos dos modelos de predição de $LogK_{oc}$ (n = 639)

Modelo	R^2	$RMSE_{tr}$	RSS_{tr}	CCC_{tr}	F
M1	0,850	0,428	116,857	0,919	3597,542
M2	0,790	0,505	162,885	0,883	2400,942
M3	0,811	0,479	146,848	0,896	2732,709
M4	0,784	0,513	167,916	0,879	2309,908
M5	0,850	0,428	116,816	0,919	3598,996
M6	0,827	0,459	134,592	0,905	3039,562
M7	0,850	0,428	116,906	0,919	3595,749

A validação interna dos modelos foi realizada pelas técnicas de validação cruzada LOO e LMO. Os dados da Tabela 3 possibilitam perceber que os valores de Q^2_{LOO} e Q^2_{LMO} para os modelos M1, M5 e M7 são os maiores do conjunto e também são bem próximos aos valores de R^2 (Tabela 2). Assim, esses modelos podem ser considerados estáveis e robustos. Os modelos M1, M5 e M7 são os que possuem os menores valores de $RMSE_{cv}$ e os maiores valores deCCC_{cv}, o que confirma que esses modelos são os melhores.

Tabela 3 Dados estatísticos da validação interna (n = 639)

Modelo	Q^2_{LOO}	$RMSE_{cv}$	$PRESS_{cv}$	CCC_{cv}	Q^2_{LMO}	R^2_{Yscr}	Q^2_{Yscr}	$RMSE\_AV_{Yscr}$
M1	0,849	0,429	117,708	0,918	0,848	0,0017	-0,0046	1,1017
M2	0,789	0,507	164,231	0,882	0,788	0,0015	-0,0048	1,1017
M3	0,809	0,482	148,194	0,895	0,808	0,0016	-0,0047	1,1017
M4	0,782	0,515	169,230	0,878	0,784	0,0015	-0,0047	1,1017
M5	0,848	0,430	117,876	0,918	0,848	0,0016	-0,0047	1,1017
M6	0,825	0,461	135,784	0,825	0,825	0,0015	-0,0048	1,1017
M7	0,848	0,429	117,797	0,918	0,848	0,0016	-0,0047	1,1017

Para descartar a possibilidade de correlação ao acaso dos modelos, os valores de R^2 e de Q^2_{LOO} devem ser maiores do que R^2_{Yscr} e Q^2_{Yscr} e o valor de $RMSE_{cv}$ menor do que $RMSE\_AV_{Yscr}$. Todos os modelos avaliados atenderam esse requisito.

Os resultados da validação externa para o conjunto de teste considerado (Tabela 4), novamente, indicaram que os melhores modelos foram os que usaram os algoritmos ALOGPs, KOWWIN e XLOGP3, a saber: M1, M5 e M7. Esses modelos apresentaram os maiores valores para R^2_{ext}, CCC_{ext} e $\overline{r^2_m}$ e os menores valores para $RMSE_{ext}$, $PRESS_{ext}$ e Δr^2_m, atendendo às recomendações da literatura (CHIRICO; GRAMATICA, 2011; 2012; ROY et al., 2012).

Tabela 4 Dados estatísticos da validação externa (n = 321)

Modelo	R^2_{ext}	$RMSE_{ext}$	$PRESS_{ext}$	CCC_{ext}	$\overline{r^2_m}$	Δr^2_m
M1	0,810	0,480	73,810	0,897	0,733	0,121
M2	0,755	0,545	95,247	0,864	0,659	0,155
M3	0,732	0,572	104,880	0,852	0,631	0,135
M4	0,776	0,522	87,299	0,877	0,686	0,150
M5	0,797	0,496	79,084	0,891	0,716	0,110
M6	0,782	0,515	85,224	0,881	0,696	0,116
M7	0,792	0,504	81,484	0,888	0,710	0,088

Considerando que a sorção no solo de compostos não iônicos ocorre por meio de interações hidrofóbicas regidas pelas forças de van der Waals, a hidrofobicidade de uma molécula, medida pelo LogP, é fundamental no processo. Assim, na impossibilidade de se utilizarem valores experimentais de LogP para estudo de fenômenos cuja hidrofobicidade é um fator importante, os valores estimados de LogP utilizados devem ser aqueles que mais se aproximam aos valores experimentais. Neste estudo, os melhores modelos foram obtidos a partir de valores de LogP calculados pelos algoritmos ALOGPs, KOWWIN e XLOGP3. O resultado está de acordo com o estudo de Reis et al. (2013), que avaliaram esses algoritmos na modelagem de $LogK_{oc}$ de pesticidas não iônicos. A boa capacidade de predição de LogP desses algoritmos e, consequentemente, a melhor qualidade dos modelos obtidos a partir deles, pode ser atribuída ao tamanho e diversidade do conjunto de dados e à abordagem de cálculo usada na calibração de cada algoritmo (MANNHOLD; Van Der WATERBEEMD, 2001).

3.2 Equivalência estatística dos modelos

A verificação da equivalência estatística entre os modelos foi realizada utilizando-se o procedimento proposto por Brownlee (1965), apresentado na seção 2.4. (página 29). Assim, se as variâncias são iguais ($F_1 < F_c$), se há paralelismo entre as retas de regressão ($|T_1| < t_c$)

e se os interceptos dos modelos são iguais ($|T_2| < t_c$) considera-se que os modelos são estatisticamente equivalentes.

Neste estudo, ao determinar que os algoritmos mais eficientes para a modelagem de $LogK_{oc}$ foram ALOGPs, KOWWIN e XLOGP3, isto é, os modelos M1, M5 e M7, comparou-se cada um desses modelos com todos os modelos considerados. Os resultados podem ser vistos nas tabelas 5, 6 e 7 e apontam que existe equivalência estatística somente entre os modelos M1, M5 e M7.

Tabela 5 Comparação do modelo M1 com os outros modelos (n1 = n2 = 639)

Modelo	F_1	$F_c = F(n_1\text{-}2,n_2\text{-}2)$	$\|T_1\|$	$t_c = t(n_1+n_2\text{-}4)$	$\|T_2\|$	$t_c = t(n_1+n_2\text{-}3)$
M2	1,44	1,14	2,867	1,962	2,388	1,962
M3	1,28	1,14	3.910	1,962	43,181	1,962
M4	1,44	1,14	4,649	1,962	0,940	1,962
M5	1,00	1,14	0,347	1,962	0,362	1,962
M6	1,17	1,14	0,892	1,962	0,090	1,962
M7	1,00	1,14	1,215	1,962	0,758	1,962

Tabela 6 Comparação do modelo M5 com os outros modelos (n1 = n2 = 639)

Modelo	F_1	$F_c = F(n_1\text{-}2,n_2\text{-}2)$	$\|T_1\|$	$t_c = t(n_1+n_2\text{-}4)$	$\|T_2\|$	$t_c = t(n_1+n_2\text{-}3)$
M1	1,00	1,14	0,347	1,962	0,362	1,962
M2	1,44	1,14	2,560	1,962	2,065	1,962
M3	1,28	1,14	3,591	1,962	0,568	1,962
M4	1,44	1,14	4,346	1,962	0,616	1,962
M6	1,17	1,14	0,561	1,962	0,436	1,962
M7	1,00	1,14	0,868	1,962	0,396	1,962

Tabela 7 Comparação do modelo M7 com os outros modelos (n1 = n2 = 639)

Modelo	F_1	$F_c = F(n_1\text{-}2,n_2\text{-}2)$	$\|T_1\|$	tc = t(n1+n2-4)	$\|T_2\|$	tc = t(n1+n2-
M1	1,00	1,14	1,215	1,962	0,758	1,962
M2	1,44	1,14	1,793	1,962	1,712	1,962
M3	1,28	1,14	2,792	1,962	0,200	1,962
M4	1,44	1,14	3,585	1,962	0,263	1,962
M5	1,00	1,14	0,868	1,962	0,396	1,962
M6	1,17	1,14	0,267	1,962	0,815	1,962

Esses resultados mostram que, na impossibilidade do acesso a um dos três algoritmos (i.e., ALOGPs, KOWWIN ou XLOGP3) para a modelagem de $LogK_{oc}$, o uso de outro será equivalente.

3.3 Modelo QSPR da média dos valores de LogP dos três melhores modelos

Após análise dos resultados obtidos, decidiu-se por construir um novo modelo para predição do $LogK_{oc}$, designado M8, a partir de valores médios de LogP (média aritmética dos valores calculados pelos algoritmos ALOGPs, KOWWIN e XLOGP3). O modelo e seus parâmetros estatísticos são apresentados na Tabela 8.

Tabela 8 Parâmetros estatísticos do modelo M8

Equação do Modelo					
$LogK_{oc}$ = 1,292 + 0,543 MEDIA					
Ajuste (n=639)		**Validação interna (n=639)**		**Validação externa (n=321)**	
R^2	0,860	Q^2_{LOO}	0,859	R^2_{ext}	0,810
$RMSE_{tr}$	0,413	$RMSE_{cv}$	0,415	$RMSE_{ext}$	0,481
RSS_{tr}	108,938	$PRESS_{cv}$	109,791	$PRESS_{ext}$	74,126
CCC_{tr}	0,925	CCC_{cv}	0,924	CCC_{ext}	0,898
F	3905,363	Q^2_{LMO}	0,859	$\overline{r_m^2}$	0,733
		R^2_{Yscr}	0,0016	Δr_m^2	0,099
		Q^2_{Yscr}	-0,0047		
		$RMSE\ AV_{Y\text{-}SCR}$	1,1017		

Os dados apresentados na Tabela 8 mostram que o modelo M8 possui qualidade estatística ligeiramente superior à dos modelos originais (M1, M5 e M7). Esse resultado pode ser atribuído ao fato de que o valor médio de LogP está mais próximo do valor real do que os valores de LogP estimados pelos algoritmos isoladamente. Assim, ao utilizar os valores médios de LogP, como os algoritmos de LogP adotaram abordagens de cálculo diferentes e foram calibrados a partir de conjuntos de compostos diferentes, a deficiência apresentada por um dado algoritmo para estimar valores de LogP, para determinadas classes de compostos, pode ser compensada pela eficiência de outro para essas mesmas classes, e vice-versa.

No que diz respeito à equivalência estatística desse modelo, em relação aos melhores modelos obtidos neste trabalho (i.e., M1, M5 e M7), verificou-se que o modelo da média é equivalente a esses três modelos (Tabela 9).

Tabela 9 Comparação do modelo M8 (média) com os 3 melhores modelos (n1 = n2 = 639)

Modelo	F_1	$F_c = F(n_1-2,n_2-2)$	$\|T_1\|$	$t_c = t(n1+n2-4)$	$\|T_2\|$	$t_c = t(n1+n2-$
M1	1,06	1,14	1,046	1,962	0,788	1,962
M5	1,06	1,14	0,693	1,962	0,420	1,962
M7	1,06	1,14	0,190	1,962	0,018	1,962

3.4 Comparação com modelos QSPR da literatura

Os principais modelos obtidos neste estudo foram comparados com modelos apresentados recentemente na literatura. Pela análise dos dados apresentados na Tabela 10, verificou-se que o melhor modelo deste estudo (M8) apresenta valores que o colocam entre o melhor modelo de Shao et al. (2014) e o modelo de Wang et al. (2015). Já quando foram considerados os modelos M1, M5 e M7, em relação ao modelo de Wang et al. (2015), verificou-se que todos eles têm melhor capacidade de predição (maiores valores de R^2_{ext} e menores valores de $RMSE_{ext}$).

Tabela 10 Comparação de parâmetros estatísticos entre os melhores modelos deste estudo e modelos anteriores

				Qualidade do ajuste			Robustez	Capacidade de predição		
Estudos	**Modelo[1]**	**K[2]**	**N[3]**	**N tr[4]**	**R^2**	**$RMSE_{tr}$**	**Q^2_{Loo}**	**N ext[5]**	**R^2_{ext}**	**$RMSE_{ext}$**
Shao et al. (2014)	LS-SVM	4	964	643	0,904	0,344	0,840	321	0,846	0,431
	GA-MLR	4	964	644	0,817	0,490	0,813	320	0,808	0,475
	LLR	4	NA	NA	0,873	0,398	0,824	NA	0,831	0,450
Wang et al. (2015)	MLR	9	824	618	0,854	0,472	0,850	206	0,761	0,558
Este estudo	M1	1	960	639	0,850	0,428	0,849	321	0,810	0,480
	M5	1	960	639	0,850	0,428	0,848	321	0,797	0,496
	M7	1	960	639	0,850	0,428	0,848	321	0,792	0,504
	M8	1	960	639	0,860	0,413	0,859	321	0,810	0,481

Notas: [1] LS-SVM = *Least Squares-Support Vector Machine*; GA-MLR = *Genetic Algorithm-Multiple Linear Regression*, LLR = *Local Lazy Regression*, MLR = *Multiple Linear Regression*; [2] K = número de descritores do modelo; [3] N = número de compostos usados; [4] N tr = número de compostos do conjunto de treinamento; [5] N ext = número de compostos do conjunto de teste; [6] NA = Não apresentado no artigo original.

Adaptado de Wang et al. (2015).

Um aspecto a favor dos modelos apresentados neste estudo é o fato de terem sido desenvolvidos com um único descritor (LogP), enquanto os modelos desenvolvidos por Shao et al. (2014) utilizaram 4 variáveis explicativas e o modelo de Wang et al. (2015) utilizou 9 variáveis. Assim, duas questões podem ser ressaltadas a partir deste fato: 1) os melhores valores de ajuste do melhor modelo de Shao et al. (2014) podem ser explicados pelo uso de um número maior de variáveis explicativas e 2) o uso de vários descritores pode aumentar a complexidade da interpretação do modelo, do ponto de vista dos mecanismos químicos envolvidos. Dessa forma, os modelos apresentados neste estudo, por terem sido desenvolvidos em função de unicamente o LogP, permitem uma explicação físico-química mais simples e são confiáveis, no que se refere à qualidade estatística e ao poder de predição dos mesmos.

4 CONCLUSÕES

Neste estudo, do conjunto de algoritmos gratuitos para cálculo de LogP considerados, concluiu-se que os melhores modelos QSPR para predizer o coeficiente de sorção do solo de compostos orgânicos não iónicos foram obtidos usando os algoritmos ALOGPs, KOWWIN e XLOGP3. Essa conclusão coincide com aquela apresentada em Reis et al. (2013) e a amplia uma vez que o conjunto de dados deste trabalho inclui, além de pesticidas, outras classes de compostos orgânicos não iônicos.

Este estudo também demonstrou a importância e a utilidade do teste de equivalência estatística proposto. O resultado do teste aplicado permitiu afirmar que os modelos obtidos dos algoritmos ALOGPs, KOWWIN e XLOGP3 são estatisticamente equivalentes.

Adicionalmente, verificou-se que os modelos apresentados neste estudo possuem qualidade estatística e capacidade de predição compatíveis à de modelos mais complexos publicados recentemente na área de QSPR.

Por fim, sugere-se que em novos estudos QSPR sejam utilizados valores de LogP, obtidos a partir da média dos valores dados pelos melhores algoritmos.

AGRADECIMENTOS

Os autores agradecem ao CNPq/MCT/Brasil pelo suporte financeiro, aos professores Miguel Angel Uribe Opazo (Estatística/UNIOESTE) e Silvia Nagib Elian (Estatística/IME/USP) pelas suas contribuições no teste de equivalência estatística e ao Grupo de Pesquisa sobre QSAR em Química Ambiental e Ecotoxicologia do Departamento de Ciências Teóricas e Aplicadas da Universidade de Insubria – Varese-Itália (DiSTA/UNINSUBRIA) por fornecer o programa QSARINS 2.2.1.

REFERÊNCIAS

ALLEN-KING, R. M.; GRATHWOHL, P.; BALL, W. P.; New modeling paradigms for the sorption of hydrophobic organic chemicals to heterogeneous carbonaceous matter in soils, sediments, and rocks. **Adv. Water Resour**., v. 25, p. 985-1016, 2002.

BROWNLEE, K. A.; **Statistical theory and methodology in science and engineering**. 2. ed. New York, NY - USA : John Wiley & Sons, 1965. 590 p.

CHENG, T.; ZHAO, Y.; LI, X.; LIN, F.; XU, Y.; ZHANG, X.; LI, Y.; WANG, R.; LAI, L.; Computation of octanol-water partition coefficients by guiding an additive model with knowledge. **J. Chem. Inf. Model**., v. 7, n. 6, p. 2140-2148, 2007.

CHIRICO N.; GRAMATICA P.; Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. **J. Chem. Inf. Model.**, v. 51, n. 9, p. 2320–2335, 2011.

CHIRICO, N.; GRAMATICA, P.; Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. **J. Chem. Inf. Model.**, v. 52, n. 8, p. 2044-2058, 2012.

CRONIN, M. T. D.; LIVINGSTONE, D. (Eds.). **Predicting chemical toxicity and fate.** Boca Raton, FL – USA: CRC Press, 2004.

DOUCETTE, W. J. Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. **Environ. Toxicol. Chem.**, v. 22, n. 8, p. 1771-1788, 2003.

GAUDIO, A. C.; ZANDONADE, E. Proposition, validation and analysis of QSAR models. **Química Nova**, São Paulo - SP, v. 24, n. 5, p. 658-671, 2001.

GAWLIK, B. M.; SOTIRIOU, N.; FEICHT, E. A.; SCHULTE-HOSTEDE, S.; KETTRUP, A. Alternatives for the determination of the soil adsorption coefficient, KOC, of non-ionicorganic compounds - a review. **Chemosphere**, v. 34, n. 12, p. 2525-2551, 1997.

GOUDARZI, N.; GOODARZI, M.; ARAUJO, M. C. U.; GALVÃO, R. K. H. QSPR modeling of soil sorption coefficients (KOC) of pesticides using SPA-ANN and SPA-MLR. **J. Agric. Food Chem.**, v. 57, n. 15, p. 7153–7158, 2009.

GRAMATICA, P.; CORRADI, M.; CONSONNI, V. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. **Chemosphere**, v. 41, n. 5, p. 763-777, 2000.

GRAMATICA, P.; GIANI, E.; PAPA, E. Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. **J. Mol. Graph. Model.**, v. 25, n. 6, p. 755-766, 2007.

GRAMATICA, P. On the development. and validation of QSAR models. *In*: REISFELD, B.; MAYENO, A. N. (Eds.).**Computational Toxicology**: New York: Human Press, 2013. Volume II. (Series: Methods in molecular biology, v. 930).

GRAMATICA, P.; CHIRICO, N.; PAPA, E.; CASSANI, S.; KOVARICH, S. QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. **J. Comp. Chem.**, v. 34, p. 2121-2132, 2013.

GRAMATICA, P.; CASSANI, S.; CHIRICO, N.; QSARINS-Chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. **J. Comp. Chem.**, v. 35, p. 1036-1044, 2014.

HANSCH, C.; LEO, A.; HOEKMAN, D. **Exploring QSAR**: hydrophobic, electronic, and steric constants. Washington, DC - USA: American Chemical Society, 1995. 348 p.

HUUSKONEN, J. Prediction of soil sorption coefficient of organic pesticides from the atom-type electrotopological state indices. **Environ. Toxicol. Chem.**, v. 22, n. 4, p. 816-820, 2003.

KIRALJ, R.; FERREIRA, M. M. C. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. **J. Braz. Chem. Soc.**, v. 20, n. 4, p. 770-787, 2009.

LIN, L. A concordance correlation coefficient to evaluate reproducibility. **Biometrics**, v. 45, n. 1, p. 255-268, 1989.

MACKAY, D.; SHIU, W. Y.; MA, K. C. LEE, S. C. Handbook of physical-chemical properties and environmental fate for organic chemicals. 2. ed. Boca Raton, FL - USA: CRC Press, 2006. 4216 p.

MANNHOLD, R.; VAN DER WATERBEEMD, H. Substructure and whole molecule approaches for calculating LogP. **J. Comput-Aided. Mol. Des.**, v. 15, n. 4, p. 337-354, 2001.

MANNHOLD, R.; PODA, G. I.; OSTERMANN, C.; TETKO, I. G. Calculation of molecular lipophilicity: state-of-the-art and comparison of LogP methods on more than 96.000 Compounds. **J. Pharm. Sci.**, v. 98, n. 3, p. 861-893. 2009.

OJHA, P. K.; MITRA, I.; DAS, R. N.; ROY, K. Further exploring r_m^2 metrics for validation of QSPR models. **Chemom. Intell. Lab. Syst.**, v. 107, n. 1, p. 194-205, 2011.

RAZZAQUE, M. M.; GRATHWOHL, P. Predicting organic carbon-water partitioning of hydrophobic organic chemicals in soils and sediments based on water solubility. **Water Research**, v. 42, n. 14, p. 3775-3780, 2008.

REIS, R. R.; SAMPAIO, S. C.; MELO, E. B. An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficient. **Water Research**, v. 53, p. 191-199, 2014.

REIS, R. R.; SAMPAIO, S. C.; MELO, E. B. The effect of different log P algorithms on the modeling of the soil sorption coefficient of nonionic pesticides. **Water Research**, v. 47, p. 5751-5759, 2013.

ROWE, P. H. Statistical methods for continuous measured endpoints in *In Silico* toxicology. *In*: CRONIN, M. T. D.; MADDEN, J. C. (eds). ***In silico* toxicology**: principles and applications, issues in toxicology. n. 7. Cambridge - UK : RSCPublishing, 2010.

ROY, K.; MITRA, I.; KAR, S.; OJHA, P.K.; DAS, R.N.; KABIR, H. Comparative studies on some metrics for external validation of QSPR models. **J. Chem. Inf. Model.**, v. 52, n. 2, p. 396-408. 2012.

RÜCKER, C.; RÜCKER, G.; MERINGER, M. Y-Randomization and its variants in QSPR/QSAR. **J. Chem. Inf. Model**., v. 47, n. 6, p. 2345–2357, 2007.

SABLJIC, A.; GÜSTEN, H.; VERHAAR, H.; HERMENS, J. QSAR modeling of soil sorption. Improvements and systematics of log Koc vs log Kow correlations. **Chemosphere**, v. 31, n. 11-12, p. 4489-4514, 1995.

SCHÜÜRMANN, G.; EBERT, R. U.; KÜHNE, R. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. **Environ. Sci. Technol.,** v. 40, n. 22, p. 7005–7011, 2006.

SHAO, Y.; LIU, J.; WANG, M.; SHI, L.; YAO, X.; GRAMATICA, P. Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods. **Atmos. Environ.**, v. 88, p. 212-218, 2014.

TAO, S.; LU, X. X.; CAO, J.; DAWSON, R. A comparison of the fragment constant and molecular connectivity indices models for normalized sorption coefficient estimation. **Water Environ. Res.**, v. 73, p. 307–313, 2001.

TETKO, I. V.; TANCHUK, V. Y.; VILLA, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. **J. Chem. Inf. Comput. Sci.**, **v.** 41, n. 5, p. 1407-1421, 2001.

TODESCHINI, R.; CONSONNI, V. **Molecular descriptors for chemoinformatics.** Weinheim - Germany : Wiley-VCH, 2009. 1257 p.

WANG, Y.; CHEN, J.; YANG, X.; LYAKURWA, F.; LI, X.; QIAO, X. *In silico* model for predicting soil organic carbon normalized sorption coefficient (KOC) of organic chemicals. **Chemosphere,** v, 119, n. 5, p. 438-444, 2015.

WEN, Y.; SU, L.M.; QIN, W. C.; FU, L.; HE, J.; ZHAO, Y. H. Linear and non-linear relationships between soil sorption and hydrophobicity: model, validation and influencing factors. **Chemosphere**, v. 86, n. 6, p. 634-640, 2012.

ARTIGO 2 – EQUIVALÊNCIA ESTATÍSTICA DE MODELOS DE PREDIÇÃO DO COEFICIENTE DE SORÇÃO DO SOLO ($LOGK_{OC}$) OBTIDOS A PARTIR DE CONJUNTOS DE TREINAMENTO DE TAMANHOS DIFERENTES[2]

1 INTRODUÇÃO

A crescente preocupação com o meio ambiente tem provocado a realização de diversos estudos que permitem predizer propriedades físico-químicas e atividades biológicas de compostos químicos pelo uso de métodos computacionais.

A modelagem das relações quantitativas estrutura-atividade (QSAR) é uma abordagem que vem ganhando espaço em vários campos da ciência. Utilizados inicialmente na área farmacéutica para a predição do comportamento de novos medicamentos, modelos QSPR na área ambiental têm sido utilizados para avaliação de riscos, predição de contaminação e tomada de decisão em questões regulatórias.

Assim, assumindo que a estrutura molecular de um composto químico contém características que definem as propriedades fisico-quimicas e biológicas desse composto, a partir do uso de modelos QSAR pode-se, por exemplo, predizer a atividade biológica de um novo composto químico que tenha uma estrutura molecular semelhante à de compostos cuja atividade biológica tenha sido avaliada experimentalmente (GRAMATICA, 2013).

---

[2] Artigo submetido à revista *Water Research* (Oxford) – ISSN: 0043-1354– Classificação A1 em Periódicos Qualis 2015 - Ciências Agrárias I.

O desenvolvimento de modelos QSPR pode ser dividido em três etapas, a saber: 1) preparação dos dados; 2) análise dos dados; 3) validação do modelo (GOLBRAIKH et al., 2003). A primeira etapa implica a escolha de um conjunto de dados experimentais sobre uma atividade biológica para um conjunto de compostos já testado; no cálculo dos descritores moleculares para esse conjunto de compostos e na determinação dos métodos estatísticos para encontrar e validar a relação entre esses dados. A segunda fase se refere à construção dos modelos QSPR que correlacionam os valores dos descritores moleculares com os valores da atividade biológica, e a terceira etapa consiste na validação do modelo, no que diz respeito à capacidade de predizer a atividade biológica de compostos ainda não estudados (GOLBRAIKH et al., 2003).

É claro que é de fundamental importância a escolha do conjunto de dados experimentais usado para o desenvolvimento do modelo QSPR, porquanto a validade estatística de um modelo QSPR depende da validade dos dados que foram utilizados para o seu desenvolvimento (TROPSHA, 2010; GRAMATICA, 2013).

Vários estudos foram feitos visando entender a relação entre a capacidade de predição de um modelo QSPR e o conjunto de treinamento utilizado para gerá-lo (FURUSJÖ et al., 2006; GOLBRAIKH et al., 2003; LEONARD; ROY, 2006; MARTIN et al., 2012; PUZYN et al., 2011; ROY; LEONARD; ROY, 2008). Alguns desses estudos sugerem que modelos QSPR baseados em conjuntos de treinamento e de teste gerados a partir do uso de métodos racionais para dividir o conjunto de dados seriam mais confiáveis do que utilizando métodos de divisão randômicos (GOLBRAIKH et al., 2003; MARTIN et al., 2012). Outros estudos apontam a necessidade da proximidade, no espaço multidimensional dos descritores, entre os pontos que representam os compostos químicos no conjunto de treinamento e os pontos do conjunto de teste (GOLBRAIKH; TROPSHA, 2002; LEONARD; ROY, 2006). Puzyn et al. (2011) investigaram de que forma o método utilizado para separar o conjunto de dados em conjunto de treinamento e conjunto de teste influência na capacidade de predição do modelo QSPR e Roy, Leonard e Roy (2008), utilizando diferentes conjuntos de dados de tamanho

moderado e várias técnicas estatísticas, avaliaram de que maneira o tamanho do conjunto de treinamento impactava a capacidade de predição do modelo, concluíndo não ser possível definir uma regra sobre isto e apontando que o tamanho ideal do conjunto de treinamento depende do conjunto de dados considerado, dos descritores e das análises estatísticas utilizadas.

Um parâmetro físico-químico que tem sido utilizado para determinar o destino final das substâncias químicas lançadas na natureza é o coeficiente de sorção do solo normalizado para o conteúdo de carbono orgânico (K_{oc}) (DOUCETTE, 2003; HUUSKONEN, 2003). Considerando o importante papel que a hidrofobicidade tem no processo de sorção no solo de substâncias não iónicas, o logaritmo do coeficiente de partição octanol/água (LogP), que descreve o comportamento hidrofóbico ou hidrofílico de um composto, tem sido usado no desenvolvimento de modelos QSPR para predizer a capacidade de sorção de um determinado composto químico ($LogK_{oc}$) (CRONIN; LIVINGSTONE, 2004; REIS; SAMPAIO; MELO, 2013; GAWLIK et al., 1997; RAZZAQUE; GRATHWOHL, 2008).

Neste trabalho, o interesse foi o de mostrar que, a partir de conjuntos de treinamento não tão grandes, podem ser desenvolvidos modelos QSPR para predição de $LogK_{oc}$ a partir de valores de LogP, que são estatísticamente equivalentes e que têm capacidade de predição similar daqueles desenvolvidos a partir de um conjunto de treinamento maior. Para tal, utilizando-se os valores experimentais do $LogK_{oc}$ para 964 compostos orgânicos não iônicos usados por Shao et al. (2014), neste trabalho, foram desenvolvidos modelos QSPR para $LogK_{oc}$ a partir de valores de LogP calculados pelo algoritmo ALOGPs, para cada composto do conjunto. O algoritmo desenvolvido se mostrou mais adequado para esses fins, comparativamente a outros algoritmos existentes (i.e., AC_logP, ALOGP, MLOGP, KOWWIN, XLOGP2 e XLOGP3), conforme resultados apresentados por Reis, Sampaio e Melo (2013), considerando pesticidas e por outro trabalho escrito pelo grupo desta pesquisa (em processo de avaliação), usando o conjunto de dados de Shao et al. (2014). Modelos foram gerados considerando-se o conjunto de treinamento total e subconjuntos do mesmo (i.e., metades,

quartos e oitavos), validados em relação à sua confiabilidade e capacidade de predição, considerando, para todos os casos, o conjunto de testes completo (321 compostos) e as recomendações feitas em Chirico e Gramatica (2011); Kiralj e Ferreira (2009) e Roy et al. (2012). e, finalmente, Verifiou-se a esxistência ou não de equivalência estatística entre os modelos obtidos a partir dos subconjuntos estudados, pois a existência dessa equivalência permitiria mostrar que modelos gerados a partir de conjuntos de treinamento menores possibilitam obter resultados equivalentes no que diz respeito à capacidade de predição dos modelos.

2 MATERIAIS E MÉTODOS

2.1 Valores experimentais de $LogK_{oc}$

Os valores experimentais do coeficiente de sorção do solo ($LogK_{oc}$) utilizados neste trabalho foram extraídos de Shao et al. (2014). O conjunto foi selecionado por ser extenso (contém 964 compostos orgânicos não iônicos) e heterogêneo, pois apresenta compostos de diferentes classes químicas. Os valores experimentais do $LogK_{oc}$ usados são apresentados no material suplementar (Tabela S1 – Apêndice B). O mesmo método foi utilizado para dividir o conjunto de dados, Y-ranking[3], o qual, segundo Puzyn et al. (2011), "produz dois conjuntos, de treinamento e de teste, que representam os dados com precisão". Esse método se baseia na ordenação da variável dependente (valores experimentais de $LogK_{oc}$), na divisão do

[3] Este procedimento é também nomeado como *activity ranking* e Z:1 em Golbraikh e Tropsha (2002) e Puzyn et al. (2011), respectivamente.

conjunto total em “caixas” de igual tamanho (3 elementos no nosso caso) e na seleção sistemática de compostos para compor o conjunto de teste e de treinamento (i.e., primeiro composto da caixa para teste, dois seguintes para treinamento).

Assim, como resultado dessa divisão, obteve-se um conjunto de treinamento formado por 643 compostos e um conjunto de validação externa, formado por 321 compostos. Quatro compostos do conjunto de treinameto usado por Shao et al. (2014) foram descartados (i.e., glyphosate, p-benzidine, ciprofloxacin e enrofloxacin), pois apresentaram resíduo maior do que 5,50. Assim, os modelos QSPR deste trabalho foram obtidos a partir de um conjunto de treinamento formado por 639 compostos (valores experimentais do $LogK_{oc}$ variando entre -0,386 e 6,469). A avaliação da capacidade de predição de todos os modelos obtidos foi realizada considerando-se o conjunto de teste gerado no processo de divisão do conjunto inicial, formado por 321 compostos (valores experimentais do $LogK_{oc}$ variam de -0,630 até 6,100). Os 960 compostos do conjunto de dados total foram classificados em 29 grupos de acordo com diferenças no grupo funcional predominante de cada composto (Tabela 1).

Tabela 1 Classificação dos 960 compostos baseados em diferenças no grupo funcional predominante

Id. do Grupo	Grupo	$N^{(1)}$	$N_{tr}^{(2)}$	$N_{ext}^{(3)}$
G1	Ácido Orgânico	67	48	19
G2	Alcano	26	16	10
G3	Alcano Halogenado	64	43	21
G4	Alceno Halogenado	13	9	4
G5	Alcenos e Alcinos	38	29	9
G6	Álcool	53	37	16
G7	Amida	16	13	3
G8	Amina	30	19	11
G9	Anilinas	44	28	16
G10	Benzeno e Alquil Benzeno	36	23	13
G11	Benzeno Halogenado	31	18	13
G12	Bifenil	41	27	14
G13	Compostos Carbonílicos	35	26	9
G14	Derivados Benzênicos	85	68	17
G15	Éster	45	31	14
G16	Éter	20	12	8
G17	Fenil Ureia	24	21	3
G18	Fenóis	66	36	30
G19	Heterociclo	10	6	4
G20	Heterociclo Aromático	35	19	16
G21	Heterociclo Poliaromático	14	6	8
G22	HPA	49	29	20
G23	Nitrila	16	10	6
G24	Nitroalcano	6	5	1
G25	Nitrobenzeno	22	10	12
G26	Organo Fosforado	19	17	2
G27	Organossulfurado	18	9	9
G28	Triazinas	7	4	3
G29	Outros Compostos	30	20	10
	Total	960	639	321

Notas:

(1) N = número de compostos do conjunto de dados;

(2) N_{tr} = número de compostos do conjunto de treinamento;

(3) N_{ext} = número de compostos do conjunto de teste.

Na Tabela S2 do material suplementar (Apêndice B) são apresentados os critérios utilizados para classificar os compostos químicos considerados neste trabalho.

O uso desse algoritmo para divisão dos dados permitiu a obtenção de conjuntos cujos compostos estão uniformemente distribuídos no espaço de dados definido pelos valores

experimentais de $LogK_{oc}$, garantindo os critérios de proximidade entre os pontos representativos dos conjuntos de treinamento, de teste e de diversidade de conjunto de treinamento sugerido por (GOLBRAIKH; TROPSHA, 2002). A Tabela S3 do material suplementar (Apêndice B) mostra a quantidade de compostos, por classe, considerados em cada conjunto de treinamento e no conjunto de teste. A Tabela S4 do material suplementar (Apêndice B) apresenta os valores mínimos e máximos do $LogK_{oc}$, por classe, para cada conjunto considerado (treinamento e teste).

2.2 Validação dos modelos QSPR

A validação de modelos QSPR é fundamental para garantir que os mesmos sejam bem ajustados, confiáveis, robustos e capazes de realizar predições confiáveis sobre novos compostos (GRAMATICA, 2013, 2007; TROPSHA; GRAMATICA; GOMBAR, 2003).

Neste trabalho, a qualidade do ajuste foi avaliada pelo cálculo do coeficiente de determinação (R^2), da raiz do erro quadrado médio ($RMSE_{tr}$) e do coeficiente de concordância da correlação para o conjunto de treinamento (CCC_{tr}). Assim, modelos com valores de R^2 maiores que 0,7, valores de $RMSE_{tr}$ baixos e o valores de CCC_{tr} maiores do que 0,85 são considerados bem ajustados (CHIRICO; GRAMATICA, 2011; 2012; GAUDIO; ZANDONADE, 2001; ROY et al., 2012). No entanto, esses parâmetros não dizem nada sobre a robustez dos modelos nem sobre a capacidade de predição interna destes. Nesse sentido, a técnica *Leave-One-Out* (LOO) foi utilizada para avaliar a confiabilidade estatística dos modelos. Para realizar esta validação, exclui-se, um a um, cada objeto do modelo; reconstrói-se o modelo sem o objeto excluído e calcula- se o valor desse objeto. Finalmente, calculam-se o coeficiente de determinação da validação LOO (Q^2_{LOO}) e a raiz do erro quadrado médio da validação cruzada ($RMSE_{cv}$). Segundo Chirico e Gramatica (2011; 2012), um modelo é estatisticamente confiável

se o valor de $RMSE_{cv}$ for o menor possível e o valor de Q^2_{LOO} for maior que 0,6. Além destes parâmetros, foi calculado o coeficiente de concordância da correlação da validação cruzada (CCC_{cv}) que, segundo esses mesmos autores, deve ser maior do que 0,85.

A estabilidade dos modelos foi avaliada pela validação *Leave-Many-Out* (LMO). Valores de Q^2_{LMO} próximos do valor de Q^2_{LOO} indicam que o modelo é robusto (KIRALJ; FERREIRA, 2009).

Adicionalmente, para descartar a possibilidade de que a relação entre as variáveis explicativa e a variável resposta tenha sido resultado do acaso, utilizou-se o teste Y-randomização (RÜCKER; RÜCKER; MERINGER, 2007).

Finalmente, foi avaliado o poder de predição de valores de $LogK_{oc}$ para novos compostos, para todos os modelos estudados, pelo cálculo do coeficiente de determinação da validação externa (R^2_{ext}), da raiz do erro quadrado médio da predição ($RMSE_{ext}$), do coeficiente de determinação da validação externa modificado (r^2_m) e do valor do coeficiente de concordância da correlação (CCC_{ext}). Para que o poder de predição de um modelo seja considerado bom o valor de R^2_{ext} deve ser maior do que 0,7, o valor de $RMSE_{ext}$ deve ser o menor possível, o valor de $\overline{r^2_m}$ e de Δr^2_m deve ser maior do que 0,65 e menor do que 0,2, respectivamente, e o valor de CCC_{ext} deve ser maior do que 0,85 (CHIRICO; GRAMATICA, 2011; 2012).

Após esses procedimentos de validação, os modelos obtidos foram comparados entre si para verificar a existência, ou não, de equivalência estatística entre eles.

2.3 Obtenção dos valores de LogP e dos modelos QSPR

Neste trabalho, foram utilizados os valores do coeficiente de partição octanol/água (LogP) de todos os compostos calculados pelo algoritmo ALOGPs. Para isso, utilizou-se o

programa ALOGPS 2.1, do *Virtual Computational Chemistry Laboratory*, disponibilizado em http://www.vcclab.org/lab/alogps/.

Os modelos para estimação do $LogK_{oc}$ em função do LogP foram obtidos pela regressão linear simples, utilizando QSARINS 2.2.1 (GRAMATICA et al., 2014; 2013). Ao todo, 15 modelos foram gerados, a saber: um modelo considerando o conjunto de treinamento total (A, n = 639), dois modelos obtidos pela divisão do conjunto total (H1 e H2, n=320 e 319), quatro modelos obtidos a partir da divisão das metades (Q1, Q2, Q3 e Q4, n=160, 160, 159 e 160) e oito modelos gerados a partir da divisão dos quartos (E1, E2, E3, E4, E5, E6, E7 e E8, n = 81, 79, 80, 80, 79, 81, 79 e 80). Isto é, trabalhou-se com conjuntos de treinamento bastante pequenos para a geração dos modelos (menos de 9% do conjunto de dados total, no caso dos oitavos, e menos de 17%, no caso dos quartos), usando o mesmo conjunto de teste para validar a capacidade de predição dos mesmos (321 compostos, o que representa 1/3 do conjunto total). Pode-se dizer que as divisões dos conjuntos foram aleatórias, porquanto elas ocorreram pelo envio de um elemento do conjunto para um subconjunto ou outro, de forma alternada.

A qualidade estatística desses modelos foi avaliada, comparados ao conjunto de treinamento total, a fim de verificar a equivalência estatística.

2.4 Teste de equivalência estatística entre os modelos

Neste estudo, para verificar se os modelos são estatisticamente equivalentes ao modelo gerado, a partir do conjunto total de treinamento, foi utilizado o procedimento descrito em Brownlee (1965). Esse teste consiste em verificar, para modelos obtidos a partir de conjuntos de dados diferentes, se as variâncias são iguais, se há paralelismo entre as retas

de regressão e se os interceptos dos modelos são iguais. Caso estas três igualdades sejam constatadas, os modelos são considerados equivalentes.

Assim, considerando dois grupos de observações (x11, y11), (x12, y12), ..., (x1n1, y1n1), n1 pares de dados, e (x21, y21), (x22, y22), ..., (x2n2, y2n2), n2 pares de dados, pode-se calcular os parâmetros dos modelos ajustados pelo método dos mínimos quadrados, de modo a obter as equações: $\hat{y}_{1} = \hat{\beta}_{01} + \hat{\beta}_{11}X$ e $\hat{y}_{2} = \hat{\beta}_{02} + \hat{\beta}_{12}X$.

Para verificar a igualdade das variâncias, as hipóteses $H_0 : \sigma_1^2 = \sigma_2^2 \; vs \; H_1 : \sigma_1^2 \neq \sigma_2^2$ são testadas.

A hipótese H_0 é aceita se $F_1 < F_c$ sendo

$$F_1 = \frac{Maior\{S_1^2, S_2^2\}}{Menor\{S_1^2, S_2^2\}} \sim F(n_1 - 2; n_2 - 2) \quad (1)$$

e $F(n_1 - 2; n_2 - 2) = F_c$ o ponto crítico da tabela F-Snedecor, a 5% de significância, com $n_1 - 2; n_2 - 2$ graus de liberdade no numerador e denominador, respectivamente.

O paralelismo entre as retas de regressão é avaliado comparando-se os coeficientes angulares. Assim $H_0 : \hat{\beta}_{11} = \hat{\beta}_{12} \; vs \; H_1 : \hat{\beta}_{11} \neq \hat{\beta}_{12}$ são testadas.

Para testar a hipótese H_0, usa-se a estatística:

$$T_1 = \frac{\hat{\beta}_{11} - \hat{\beta}_{12}}{\left[Var(\hat{\beta}_{11} - \hat{\beta}_{12})\right]^{1/2}} \sim t(n_1 + n_2 - 4) \quad (2)$$

assim, o paralelismo entre as retas é verificado quando $|T_1| < t_c$. Sendo $t(n_1 + n_2 - 4) = t_c$ o valor crítico da tabela t-Student bicaudal com $n_1 + n_2 - 4$ graus de liberdade e nível de 5% de significância.

Finalmente, a hipótese apropriada para verificar a igualdade dos interceptos é $H_0 : \hat{\beta}_{01} = \hat{\beta}_{02} \; vs \; H_1 : \hat{\beta}_{01} \neq \hat{\beta}_{02}$.

A estatística do teste, sob H_0, é

$$T_2 = \frac{\hat{\beta}_{01} - \hat{\beta}_{02}}{\left[Var(\hat{\beta}_{01} - \hat{\beta}_{02})\right]^{1/2}} \sim t(n_1 + n_2 - 3) \quad (3)$$

e $t(n_1 + n_2 - 3) = t_c$ é o valor crítico da tabela t-Student bicaudal com $n_1 + n_2 - 3$ graus de liberdade, ao nível de 5% de significância. Se $|T_2| < t_c$, aceita-se a hipótese nula podendo concluir que os interceptos são iguais ao nível de 5% de significância.

Assim, ao verificar as três hipóteses nulas apresentadas, conclui-se que os modelos são estatisticamente equivalentes. Uma descrição mais detalhada deste procedimento pode ser encontrada em Brownlee (1965).

3 RESULTADOS E DISCUSSÃO

3.1 Modelos QSPR de predição de LogKoc

A Tabela 2 estão apresentados os modelos de estimação do $LogK_{oc}$, calculados pelo QSARINS, considerando o algoritmo ALOGPS para cálculo do LogP e utilizando conjuntos de treinamento de tamanhos diferentes (coluna intitulada N).

Tabela 2 Modelos de predição de $LogK_{oc}$ para os conjuntos de treinamento estudados

		Exp $LogK_{oc}$		
Modelo	**N(1)**	**Mínimo**	**Máximo**	**Equação**
A	639	-0,386	6,469	$LogK_{oc}$ = 1,322 + 0,530 ALOGPs
H1	319	-0,386	6,431	$LogK_{oc}$ = 1,323 + 0,528 ALOGPs
H2	320	0,267	6,469	$LogK_{oc}$ = 1,320 + 0,532 ALOGPs
Q1	160	-0,386	5,816	$LogK_{oc}$ = 1,332 + 0,527 ALOGPs
Q2	160	0,267	6,469	$LogK_{oc}$ = 1,283 + 0,546 ALOGPs
Q3	159	0,458	6,431	$LogK_{oc}$ = 1,314 + 0,530 ALOGPs
Q4	160	0,441	6,100	$LogK_{oc}$ = 1,360 + 0,517 ALOGPs
E1	81	-0,386	5,816	$LogK_{oc}$ = 1,230 + 0,569 ALOGPs
E2	79	0,267	5,854	$LogK_{oc}$ = 1,320 + 0,520 ALOGPs
E3	80	0,556	6,431	$LogK_{oc}$ = 1,340 + 0,514 ALOGPs
E4	80	0,441	6,100	$LogK_{oc}$ = 1,289 + 0,543 ALOGPs
E5	79	-0,282	5,370	$LogK_{oc}$ = 1,409 + 0,486 ALOGPs
E6	81	0,630	6,469	$LogK_{oc}$ = 1,251 + 0,569 ALOGPs
E7	79	0,458	6,083	$LogK_{oc}$ = 1,302 + 0,541 ALOGPs
E8	80	0,958	5,277	$LogK_{oc}$ = 1,443 + 0,489 ALOGPs

Nota: (1) N=número de compostos do conjunto de treinamento**.**

Na Tabela 3 são mostrados os valores obtidos para os parâmetros estatísticos usados para verificar o ajuste dos modelos, assim como a robustez e a capacidade de predição interna dos mesmos.

Todos os modelos apresentaram valores do coeficiente de determinação (R^2) maiores que 0,7 e CCC_{tr} maiores do que 0,85, no entanto, verificou-se que os modelos que apresentam

os melhores ajustes, maior valor de R^2, são H1, Q1 e E1, gerados a partir das primeiras metades, quartos e oitavos do conjunto total de treinamento, respectivamente.

Adicionalmente, observou-se que esses três modelos são os que possuem os menores valores de $RMSE_{tr}$ e os maiores valores de CCC_{tr}, o que mostra que são os que possuem menor erro e menor diferença entre os dados experimentais e os preditos.

Pelos dados da Tabela 3, pode-se constatar, ainda, que os valores de Q^2_{LOO} e Q^2_{LMO} para os modelos H1, Q1 e E1 são os maiores do conjunto, sendo também bem próximos aos valores de R^2 do respectivo modelo. Dessa maneira, esses modelos podem ser considerados estáveis e robustos. Observa-se, ainda, que esses modelos são os que possuem os menores valores de $RMSE_{cv}$ e os maiores valores de CCC_{cv}, o que confirma que são os melhores.

Tabela 3 Parâmetros estatísticos do ajuste e da validação interna

Modelo	R^2	$RMSE_{tr}$	CCC_{tr}	Q^2_{LOO}	Q^2_{LMO}	$RMSE_{cv}$	CCC_{cv}	R^2_{Yscr}	Q^2_{Yscr}	$RMSE\_AV_{Ysc}$
A	0,850	0,428	0,919	0,849	0,848	0,429	0,918	0,0016	-0,0047	1,1017
H1	0,855	0,428	0,922	0,853	0,852	0,431	0,921	0,0032	-0,0094	1,1219
H2	0,844	0,427	0,915	0,841	0,840	0,431	0,914	0,0031	-0,0095	1,0792
Q1	0,868	0,415	0,930	0,864	0,862	0,422	0,927	0,0062	-0,0192	1,1411
Q2	0,849	0,434	0,919	0,845	0,842	0,441	0,916	0,0064	-0,0191	1,1153
Q3	0,840	0,440	0,913	0,836	0,836	0,446	0,911	0,0061	-0,0195	1,0973
Q4	0,838	0,419	0,912	0,834	0,831	0,425	0,910	0,0064	-0,0189	1,0379
E1	0,896	0,382	0,945	0,892	0,891	0,389	0,943	0,0120	-0,0390	1,1737
E2	0,855	0,396	0,922	0,844	0,844	0,411	0,916	0,0128	-0,0401	1,0347
E3	0,813	0,418	0,897	0,796	0,791	0,436	0,888	0,0127	-0,0394	0,9590
E4	0,839	0,444	0,913	0,831	0,825	0,455	0,908	0,0127	-0,0391	1,1014
E5	0,843	0,431	0,915	0,830	0,828	0,449	0,908	0,0125	-0,0400	1,0825
E6	0,849	0,463	0,918	0,841	0,842	0,475	0,914	0,0122	-0,0389	1,1815
E7	0,858	0,460	0,924	0,852	0,852	0,469	0,920	0,0129	-0,0394	1,2116
E8	0,841	0,385	0,914	0,830	0,827	0,398	0,908	0,0126	-0,0392	0,9595

Para avaliar a possibilidade de correlação ao acaso dos modelos estudados, foi utilizada a técnica Y-*scrambling*, sendo que esse fenômeno é descartado quando os valores de R^2 e de Q^2_{LOO} são maiores do que R^2_{Yscr} e Q^2_{Yscr} e o valor de $RMSE_{cv}$ é menor do que $RMSE\_AV_{Yscr}$. Assim, todos os modelos avaliados atenderam esse requisito.

Os resultados da validação externa dos modelos (Tabela 4) indicam que todos eles tem um bom poder de predição, porquanto todos apresentaram valores de R^2_{ext} maiores do que 0,7, de CCC_{ext} maiores de 0,85, de $\overline{r_m^2}$ maiores de 0,65 e de Δr_m^2 menores de 0,2, atendendo, portanto, às recomendações da literatura (CHIRICO; GRAMATICA, 2011; 2012; ROY et al., 2012). Comparando os valores da validação externa para cada tipo de subconjunto verifica-se que modelos gerados a partir das metades (H1 e H2) têm praticamente a mesma capacidade de predição; no caso dos quartos (Q1, Q2, Q3 e Q4) o modelo gerado a partir de Q2 apresenta capacidade de predição levemente melhor do que os outros, sendo que o mesmo acontece com os modelos E1, E4 e E6 no caso dos oitavos (E1 a E8).

Ainda considerando os resultados da validação externa, o fato de todos os modelos terem apresentados igual valor de R^2_{ext} (i.e., 0,810) é um indício da equivalência dos modelos.

Tabela 4 Parâmetros estatísticos da validação externa

Modelo	R^2_{ext}	$RMSE_{ext}$	CCC_{ext}	$\overline{r_m^2}$	Δr_m^2
A	0,810	0,480	0,897	0,733	0,121
H1	0,810	0,479	0,897	0,733	0,125
H2	0,810	0,480	0,897	0,733	0,118
Q1	0,810	0,479	0,897	0,732	0,132
Q2	0,810	0,482	0,899	0,734	0,077
Q3	0,810	0,480	0,897	0,733	0,116
Q4	0,810	0,479	0,895	0,729	0,158
E1	0,810	0,490	0,900	0,735	0,018
E2	0,810	0,479	0,896	0,732	0,134
E3	0,810	0,479	0,895	0,731	0,155
E4	0,810	0,482	0,899	0,734	0,084
E5	0,810	0,482	0,888	0,690	0,168
E6	0,810	0,491	0,900	0,735	0,029
E7	0,810	0,481	0,898	0,734	0,095
E8	0,810	0,482	0,888	0,682	0,170

3.2 Equivalência estatística dos modelos

A verificação da equivalência estatística entre os modelos foi realizada utilizando o procedimento proposto por Brownlee (1965), apresentado na seção 2.4 (página 29). Assim,

se as variâncias são iguais ($F_1 < F_c$), se há paralelismo entre as retas de regressão ($|T_1| < t_c$) e se os interceptos dos modelos são iguais ($|T_2| < t_c$) conclui-se que os modelos são estatisticamente equivalentes.

Neste estudo, comparou-se o modelo gerado a partir do conjunto total de treinamento (A, n = 639) com todos os modelos considerados, gerados a partir de subconjuntos do conjunto total, isto é, metades, quartos e oitavos. Os respectivos resultados podem ser vistos na Tabela 5 e mostram que existe equivalência estatística entre o modelo A, gerado a partir do conjunto de treinamento total (n = 639) e de todos os outros modelos já que todas as condições necessárias estavam cumpridas.

Tabela 5 Comparação do modelo A com os outros modelos

Modelo	F_1	Fc = F(n1-2,n2-2)	$\|T_1\|$	tc = t(n1+n2-2)	$\|T_2\|$	tc = t(n1+n2-3)
H1	1,00	1,18	0,116	1,962	0,028	1,962
H2	1,00	1,18	0,122	1,962	0,036	1,962
Q1	1,06	1,24	0,176	1,963	0,164	1,963
Q2	1,06	1,24	0,808	1,963	0,633	1,963
Q3	1,11	1,24	0,012	1,963	0,131	1,963
Q4	1,00	1,24	0,647	1,963	0,606	1,963
E1	1,20	1,34	1,539	1,963	1,124	1,963
E2	1,12	1,35	0,359	1,963	0,022	1,963
E3	1,01	1,35	0,550	1,963	0,196	1,963
E4	1,12	1,35	0,480	1,963	0,406	1,963
E5	1,06	1,35	1,756	1,963	1,172	1,963
E6	1,22	1,34	1,491	1,963	0,862	1,963
E7	1,22	1,35	0,426	1,963	0,250	1,963
E8	1,18	1,35	1,509	1,963	1,390	1,963

Os resultados sugerem que, considerando-se o conjunto de dados usado neste estudo (SHAO et al., 2014), a quantidade de compostos no conjunto de treinamento para a modelagem de $LogK_{oc}$, a partir de valores de LogP calculados pelo algoritmo ALOGPs, não precisa ser tão grande, pois o uso de conjuntos com até 8 vezes menos compostos seria equivalente do ponto de vista estatístico. Essa constatação pode ainda ser confirmada observando-se os dados apresentados nas tabelas 3 e 4, nas quais se verifica que todos os

modelos têm bom ajuste, são confiáveis e robustos e têm igual poder de predição (R^2_{ext}=0,810).

3.3 Comparação com modelos QSPR da literatura

Os modelos obtidos neste estudo foram comparados com outros apresentados recentemente na literatura. Ao analisar os dados na Tabela 6, verificou-se que todos os modelos deste estudo têm capacidade de predição similar à apresentada pelo modelo de Shao et al. (2014), desenvolvido mediante GA-MLR. No entanto, todos os modelos (exceto o E3) apresentam valores de R^2, $RMSE_{tr}$ e Q^2_{LOO} que indicam que os modelos deste estudo estão melhor ajustados e são mais robustos do que o GA-MLR obtido por Shao et al. (2014). Quando os modelos desenvolvidos na pesquisa são comparados ao modelo de Wang et al. (2015), verifica-se que todos eles têm capacidade de predição melhor (maiores valores de R^2_{ext} e menores valores de $RMSE_{ext}$). Destaca-se, no entanto, os resultados apresentados pelo modelo E1, que mostrou ajuste similar ao melhor modelo de Shao et al. (2014), desenvolvido por LS-SVM, robustez ligeiramente melhor e predição ligeiramente menor.

Tabela 6 Comparação de parâmetros estatísticos entre os modelos de este estudo e modelos anteriores

Estudo	Modelo[1]	K[2]	N[3]	Qualidade do ajuste			Robustez	Capacidade de predição		
				N_{tr}	R^2	$RMSE_{tr}$	Q^2_{loo}	N_{ext}	R^2_{ext}	$RMSE_{ext}$
Shao et al (2014)	LS-SVM	4	964	643	0,904	0,344	0,840	321	0,846	0,431
	GA-MLR	4	964	644	0,817	0,490	0,813	320	0,808	0,475
	LLR	4	NA	NA	0,873	0,398	0,824	NA	0,831	0,450
Wang et al (2015)	MLR	9	824	618	0,854	0,472	0,850	206	0,761	0,558
Este estudo	A	1	960	639	0,850	0,428	0,849	321	0,810	0,480
	H1	1	640	319	0,855	0,428	0,853	321	0,810	0,479
	H2	1	641	320	0,844	0,427	0,841	321	0,810	0,480
	Q1	1	481	160	0,869	0,415	0,864	321	0,810	0,479
	Q2	1	481	160	0,849	0,434	0,845	321	0,810	0,482
	Q3	1	480	159	0,840	0,440	0,836	321	0,810	0,480
	Q4	1	481	160	0,838	0,419	0,834	321	0,810	0,479
	E1	1	402	81	0,896	0,382	0,892	321	0,810	0,490
	E2	1	400	79	0,855	0,396	0,844	321	0,810	0,479
	E3	1	401	80	0,813	0,418	0,796	321	0,810	0,479
	E4	1	401	80	0,839	0,444	0,831	321	0,810	0,482
	E5	1	400	79	0,843	0,431	0,830	321	0,810	0,482
	E6	1	402	81	0,849	0,463	0,841	321	0,810	0,491
	E7	1	400	79	0,858	0,460	0,852	321	0,810	0,481
	E8	1	401	80	0,841	0,385	0,830	321	0,810	0,482

Notas: [1] LS-SVM = *Least Squares-Support Vector Machine*, GA-MLR = *Genetic Algorithm-Multiple Linear Regression*, LLR = *Local Lazy Regression*, MLR = *Multiple Linear Regression*; [2] K = número de descritores do modelo; [3] N = número de compostos usados; [4] N tr = número de compostos do conjunto de treinamento; [5] N ext = número de compostos do conjunto de teste [6] NA = Não apresentado no artigo original.

Adaptado de Wang et al. (2015).

4 CONCLUSÕES

Neste trabalho, verificou-se que, a partir de conjuntos de treinamento não tão grandes, modelos QSPR estatisticamente equivalentes podem ser desenvolvidos e que esses modelos têm capacidade de predição similar aos criados a partir de um conjunto de treinamento maior. Foram gerados modelos considerando o conjunto de treinamento total e subconjuntos do mesmo (i.e., metades, quartos e oitavos); validados em relação à sua confiabilidade e capacidade de predição e considerando, para todos os casos, o conjunto de testes completo (321 compostos). Todos os modelos obtiveram bons resultados, quando validados conforme as recomendações feitas em Chirico e Gramatica (2011), Kiralj e Ferreira (2009) e Roy et al. (2012).

O trabalho desenvolvido demostrou a importância do teste de equivalência estatística proposto por Brownlee (1965), porquanto permitiu afirmar que, seguindo os procedimentos adotados neste estudo, os modelos obtidos com subconjuntos do conjunto de treinamento são estatisticamente equivalentes.

AGRADECIMENTOS

Os autores agradecem ao CNPq/MCT/Brasil pelo suporte financeiro, aos professores Miguel Angel Uribe Opazo (Estatística/UNIOESTE) e Silvia Nagib Elian (Estatística/IME/USP) pelas suas contribuições no teste de equivalência estatística e ao Grupo de Pesquisa sobre QSAR em Química Ambiental e Ecotoxicologia do Departamento de Ciências Teóricas e Aplicadas da Universidade de Insubria – Varese-Itália (DiSTA/UNINSUBRIA) por fornecer o programa QSARINS 2.2.1.

REFERÊNCIAS

BROWNLEE, K. A. **Statistical theory and methodology in science and engineering**. 2. ed. New York, NY - USA: John Wiley & Sons, 1965. 590 p.

CHIRICO N.; GRAMATICA P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. **J. Chem. Inf. Model.**, v. 51, n. 9, p. 2320–2335, 2011.

CHIRICO, N.; GRAMATICA, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. **J. Chem. Inf. Model.**, v. 52, n. 8, p. 2044-2058, 2012.

CRONIN, M. T. D.; LIVINGSTONE, D. (Eds.). **Predicting chemical toxicity and fate.** Boca Raton, FL – USA: CRC Press, 2004.

DOUCETTE, W. J. Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals (Annual review). **Environ. Toxicol. Chem.**, v. 22, p. 1771-1788, 2003.

FURUSJÖ, E.; SVENSON, A.; RAHMBERG, M.; ANDERSSON, M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. **Chemosphere**, v. 63, p. 99-108, 2006.

GAUDIO, A. C.; ZANDONADE, E. Proposition, validation and analysis of QSAR models. **Química Nova**, São Paulo - SP, SBQ 24, p. 658–671, 2001.

GAWLIK, B. M.; SOTIRIOU, N.; FEICHT, E. A. SCHULTE-HOSTEDE, S.; KETTRUP, A. Alternatives for the determination of the soil adsorption coefficient, koc, of non-inorganic compounds - a review. **Chemosphere**, v. 34, p. 2525–2551, 1997.

GOLBRAIKH, A.; SHEN, M.; XIAO, Z.; XIAO, Y.; LEE, K. Rational selection of training and test sets for the development of validated QSAR models. **J. Comput. Aided. Mol. Des.**, 2003. v. 17, p. 241–253, 2003.

GOLBRAIKH, A.; TROPSHA, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. **J. Comput. Aided. Mol. Des.**, v. 16, p. 357–369, 2002.

GRAMATICA, P. On the development. and validation of QSAR models. *In*: REISFELD, B.; MAYENO, A. N. (Eds.).**Computational Toxicology**: New York: Human Press, 2013. Volume II. (Series: Methods in molecular biology, v. 930).

GRAMATICA, P. Principles Of QSAR models validation: internal and external. **QSAR Comb. Sci.**, v. 26, p. 694–701, 2007.

GRAMATICA, P.; CASSANI, S.; CHIRICO, N.; QSARINS-Chem: insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. **J. Comput. Chem.**, v. 35, p. 1036–1044, 2014.

GRAMATICA, P.; CHIRICO, N.; PAPA, E.; CASSANI, S.; KOVARICH, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. **J. Comput. Chem.**, v. 34, p. 2121–2132, 2013.

HUUSKONEN, J. Prediction of soil sorption coefficient of organic pesticides from the atom-type electrotopological state indices. **Environ. Toxicol. Chem.**, v. 22, p. 816–820, 2003.

KIRALJ, R.; FERREIRA, M. M. C. Basic validation procedures for regression models in qsar and qspr studies: theory and application. **J. Braz. Chem. Soc.**, v. 20, p. 770–787, 2009.

LEONARD, J. T.; ROY, K. On selection of training and test sets for the development of predictive QSAR models. **QSAR Comb. Sci.**, v. 25, p. 235–251. 2006.

MARTIN, T. M.; HARTEN, P.; YOUNG, D. M.; MURATOV, E. N.; GOLBRAIKH, A.; ZHU, H.; TROPSHA, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? **J. Chem. Inf. Model.**, v. 52, p. 2570–2578, 2012.

PUZYN, T.; MOSTRAG-SZLICHTYNG, A.; GAJEWICZ, A.; SKRZYŃSKI, M.; WORTH, A. P. Investigating the influence of data splitting on the predictive ability of QSAR/QSPR models. **Struct. Chem.,** v. 22, 795–804, 2011.

RAZZAQUE, M. M.; GRATHWOHL, P. Predicting organic carbon-water partitioning of hydrophobic organic chemicals in soils and sediments based on water solubility. **Water Research**, v. 42, p. 3775–3780, 2008.

REIS, R. R.; SAMPAIO, S. C.; MELO, E. B.; The effect of different logp algorithms on the modeling of the soil sorption coefficient of nonionic pesticides. **Water Research**, v. 47, p. 5751-5759, 2013.

ROY, K.; MITRA, I.; KAR, S.; OJHA, P. K.; DAS, R. N.; KABIR, H. Comparative studies on some metrics for external validation of QSPR models. **J. Chem. Inf. Model.**, v. 52, 396–408, 2012.

ROY, P. P.; LEONARD, J. T.; ROY, K. Exploring the impact of size of training sets for the development of predictive QSAR models. **Chemom. Intell. Lab. Syst.** v. 90, p. 31–42, 2008.

RÜCKER, C.; RÜCKER, G.; MERINGER, M. Y-Randomization and its variants in QSPR/QSAR. **J. Chem. Inf. Model**., v. 47, n. 6, p. 2345–2357, 2007.

SHAO, Y.; LIU, J.; WANG, M.; SHI, L.; YAO, X.; GRAMATICA, P. Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods. **Atmos. Environ.**, v. 88, p. 212–218, 2014.

TROPSHA, A. Best practices for QSAR Model development, validation, and exploitation. **Mol. Inform.**, v. 29, p. 476–488, 2010.

TROPSHA, A.; GRAMATICA, P.; GOMBAR, V. K. The Importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. **Qsar Comb. Sci.**, v. 22, p. 69–77, 2003.

WANG, Y.; CHEN, J.; YANG, X.; LYAKURWA, F.; LI, X.; QIAO, X. *In silico* model for predicting soil organic carbon normalized sorption coefficient (Koc) of organic chemicals. **Chemosphere**, v. 119, p. 438-444, 2015.

CONSIDERAÇÕES FINAIS

Neste estudo foram considerados diversos algoritmos gratuitos para cálculo de LogP, e se concluiu que os melhores modelos QSPR para predizer o coeficiente de sorção do solo de compostos orgânicos não iónicos foram obtidos usando os algoritmos ALOGPs, KOWWIN e XLOGP3.

Este estudo também demonstrou a importância e utilidade do teste de equivalência estatística proposto. O teste permitiu afirmar que os modelos obtidos dos algoritmos ALOGPs, KOWWIN ou XLOGP3 são estatisticamente equivalentes, significando que, na impossibilidade de obter valores de LogP através de um dos algoritmos, valores obtidos por outro deles podem ser usados. No entanto, quando possível, sugere-se que nos estudos QSPR sejam utilizados valores de LogP obtidos a partir da média dos valores dados pelos três melhores algoritmos.

Adicionalmente, verificou-se que os modelos apresentados neste estudo possuem qualidade estatística e capacidade de predição compatíveis a de modelos mais complexos publicados recentemente na área de QSPR.

Foi mostrado ainda que, a partir de conjuntos de treinamento não tão grandes, modelos QSPR estatisticamente equivalentes podem ser desenvolvidos e que estes modelos têm capacidade de predição similar daqueles criados a partir de um conjunto de treinamento maior. Para isto, modelos foram gerados considerando valores de LogP do conjunto de treinamento total gerados com o algoritmo ALOGPs e também com subconjuntos do mesmo (i.e., metades, quartos e oitavos).

Por fim, este estudo mostrou a importância do uso do teste de equivalência estatística proposto por Brownlee (1965) já que foi verificado que, seguindo os procedimentos adotados neste estudo, os modelos obtidos com subconjuntos do conjunto de treinamento são estatisticamente equivalentes.

APÊNDICES

APÊNDICE A TABELAS DO ARTIGO 1

Tabela 1 Nomes, números CAS, SMILES, valores experimentais de log K_{oc} e valores de log P para os compostos dos conjuntos de treinamento e de teste

Continua

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
1	training	1	bromotrifluoromethane	75-63-8	FC(F)(F)Br	2.389	1.55	3.12	2.83	1.82	1.59	1.97	2.20
2	training	2	carbon tetrabromide	558-13-4	BrC(Br)(Br)Br	3.237	3.30	3.35	4.02	2.96	2.80	4.22	3.44
3	training	3	chlorotrifluoromethane	75-72-9	FC(F)(F)Cl	2.275	1.80	3.07	2.72	1.60	1.50	1.63	2.03
4	training	4	dichlorodifluoromethane	75-71-8	FC(F)(Cl)Cl	2.552	2.06	3.25	3.01	1.82	1.82	1.91	2.27
5	training	5	trichlorofluoromethane	75-69-4	FC(Cl)(Cl)Cl	2.753	2.25	3.08	3.30	2.03	2.13	2.32	2.52
6	test	1	carbon tetrachloride	56-23-5	C(Cl)(Cl)(Cl)Cl	2.270	2.64	3.15	3.59	2.23	2.44	2.86	2.77
7	training	6	carbon tetrafluoride	75-73-0	FC(F)(F)F	2.019	1.75	1.65	2.44	1.36	1.19	1.49	1.78
8	training	7	bromoform	75-25-2	BrC(Br)Br	2.672	2.50	4.37	2.23	2.42	1.79	3.09	2.77
9	test	2	chlorodifluoromethane	75-45-6	FC(F)Cl	1.965	0.98	2.38	1.22	1.36	0.89	1.39	1.77
10	training	8	dichlorofluoromethane	75-43-4	FC(Cl)Cl	2.220	1.28	3.12	1.42	1.60	1.21	1.66	2.01
11	test	3	chloroform	67-66-3	C(Cl)(Cl)Cl	1.650	1.67	4.22	1.62	1.82	1.52	2.07	2.26
12	test	4	fluoroform	75-46-7	FC(F)F	1.725	0.97	0.95	1.03	1.12	0.58	1.25	1.52
13	training	9	bromochloromethane	74-97-5	BrCCl	2.144	1.27	2.00	1.28	1.60	1.43	1.86	1.67
14	training	10	dibromomethane	74-95-3	BrCBr	2.628	1.48	2.05	1.50	1.82	1.52	2.20	1.84
15	training	11	chlorofluoromethane	593-70-4	FCCl	1.654	0.62	1.28	0.77	1.12	1.03	1.25	1.26
16	training	12	dichloromethane	75-09-2	ClCCl	2.057	1.12	1.95	1.07	1.36	1.34	1.52	1.50
17	training	13	difluoromethane	75-10-5	FCF	1.486	0.29	0.81	0.48	0.85	0.71	1.11	1.01
18	training	14	diiodomethane	75-11-6	C(I)I	2.737	2.25	2.73	2.30	2.23	2.35	2.60	2.26
19	training	15	formaldehyde	50-00-0	O=C	1.567	-0.68	-0.25	-0.23	-0.96	0.35	0.02	1.24
20	training	16	formic acid	64-18-6	O=CO	1.083	-0.47	-0.25	-0.28	-1.03	-0.46	-0.32	-0.20
21	training	17	methyl bromide	74-83-9	BrC	0.790	0.68	1.10	0.99	1.12	1.18	1.38	0.99
22	training	18	methyl chloride	74-87-3	ClC	1.872	0.67	1.05	0.85	0.85	1.09	1.04	0.82
23	test	5	methyl fluoride	593-53-3	FC	1.654	0.41	0.72	0.58	0.55	0.77	0.90	0.57
24	test	6	methyl iodide	74-88-4	CI	1.040	1.20	1.44	1.52	1.36	1.59	1.58	1.51
25	training	19	formamide	75-12-7	O=CN	0.556	-1.53	-0.79	-0.88	-1.43	-1.61	-1.04	-0.85
26	training	20	nitromethane	75-52-5	N(=O)(=O)C	1.197	-0.17	-0.04	0.29	-0.43	-0.04	0.35	0.09
27	training	21	methane	74-82-8	C	1.970	-1.32	0.50	1.38	1.12	0.78	0.00	0.65
28	training	22	methyl alcohol	67-56-1	OC	0.974	-1.38	-0.01	-0.36	-0.81	-0.63	-0.50	-0.46
29	test	7	methylamine	74-89-5	NC	1.067	-1.06	-0.55	-0.65	-0.81	-0.64	-0.57	-0.71
30	test	8	carbon disulfide	75-15-0	C(=S)=S	2.541	2.25	1.54	1.65	-0.05	1.94	1.78	2.08
31	test	9	1,2-dichloro-1,1,2,2-tetrafluoroethane	76-14-2	FC(F)(C(F)(F)Cl)Cl	2.911	2.57	2.81	2.59	2.60	2.78	2.51	2.82

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
32	training	23	1,1,2-trichloro-1,2,2-trifluoroethane	76-13-1	FC(F)(C(F)(Cl)Cl)Cl	3.096	3.03	4.27	2.79	2.78	3.09	2.79	3.16
33	training	24	tetrachloroethylene	127-18-4	C(=C(Cl)Cl)(Cl)Cl	2.310	3.13	4.26	2.43	2.46	2.97	3.03	3.40
34	training	25	hexachloroethane	67-72-1	C(C(Cl)(Cl)Cl)(Cl)(Cl)Cl	3.553	3.93	4.33	3.38	3.29	4.03	3.88	4.14
35	training	26	hexafluoroethane	76-16-4	FC(F)(F)C(F)(F)F	2.465	2.46	2.34	2.20	2.23	2.15	2.24	2.00
36	test	10	halothane	151-67-7	FC(F)(F)C(Cl)Br	2.628	2.50	2.38	2.51	2.60	2.26	2.74	2.30
37	training	27	trichloroethylene	79-01-6	C(=CCl)(Cl)Cl	2.150	2.45	3.11	1.74	2.08	2.47	2.64	2.61
38	training	28	trichloroacetaldehyde	75-87-6	O=CC(Cl)(Cl)Cl	1.916	1.38	1.91	1.50	1.06	1.19	1.44	1.58
39	test	11	pentachloroethane	76-01-7	C(C(Cl)Cl)(Cl)(Cl)Cl	2.949	3.21	3.32	2.89	2.96	3.11	3.23	3.22
40	test	12	acetylene	74-86-2	C#C	1.578	-0.03	-0.18	1.98	0.70	0.50	0.42	0.39
41	training	29	1,1-dichloroethylene	75-35-4	C(=C)(Cl)Cl	2.536	1.97	2.70	1.69	1.67	2.12	1.93	2.32
42	test	13	cis-1,2-dichloroethylene	156-59-2	C(=CCl)Cl	2.389	1.85	1.96	1.06	1.67	1.98	2.26	1.86
43	training	30	trans-1,2-dichloroethylene	156-60-5	C(=CCl)Cl	2.427	1.85	1.96	1.06	1.67	1.98	2.26	1.86
44	training	31	dichloroacetic acid	79-43-6	O=C(O)C(Cl)Cl	1.877	0.99	0.49	0.96	0.59	0.52	0.57	0.92
45	training	32	2,2,2-trichloroacetamide	594-65-0	O=C(N)C(Cl)(Cl)Cl	1.943	0.98	0.96	0.85	0.59	0.83	0.30	1.04
46	training	33	1,1,2,2-tetrachloroethane	79-34-5	C(C(Cl)Cl)(Cl)Cl	2.677	2.57	2.31	2.40	2.60	2.19	2.58	2.39
47	training	34	1,1-difluoroethylene	75-38-7	FC(F)=C	2.052	1.56	1.19	1.11	1.22	1.24	0.87	1.26
48	test	14	trifluoroacetamide	354-38-1	O=C(N)C(F)(F)F	1.442	0.08	-0.03	0.26	-0.04	-0.11	-0.52	0.12
49	training	35	vinyl bromide	593-60-2	BrC=C	2.231	1.19	1.60	1.09	1.45	1.52	1.72	1.54
50	training	36	bromoacetic acid	79-08-3	O=C(O)CBr	1.600	0.53	0.25	0.66	0.37	0.43	0.50	0.41
51	training	37	vinyl chloride	75-01-4	C(=C)Cl	2.128	1.43	1.55	1.00	1.22	1.62	1.55	1.48
52	training	38	chloroacetic acid	79-11-8	O=C(O)CCl	1.497	0.18	0.20	0.51	0.13	0.34	0.16	0.22
53	training	39	1,1,1-trichloroethane	71-55-6	C(Cl)(Cl)(Cl)C	2.010	2.45	2.86	2.03	2.23	2.68	2.47	2.44
54	training	40	1,1,2-trichloroethane	79-00-5	ClCC(Cl)Cl	1.800	2.02	2.02	1.96	2.23	2.01	2.16	1.89
55	test	15	2,2,2-trichloroethanol	115-20-8	OCC(Cl)(Cl)Cl	2.111	1.23	1.90	1.24	1.21	1.21	1.34	1.42
56	test	16	2,2,2-trifluoroethanol	75-89-8	FC(F)(F)CO	1.600	0.61	0.91	0.65	0.58	0.27	0.51	0.41
57	test	17	acetonitrile	75-05-8	C(#N)C	1.192	-0.04	0.93	0.05	-0.32	-0.15	0.03	-0.02
58	test	18	ethylene	74-85-1	C=C	1.992	0.90	1.15	0.95	0.70	1.27	1.26	1.20
59	training	41	1,2-dibromoethane	106-93-4	BrCCBr	2.443	2.08	1.84	1.80	2.23	2.01	2.42	1.96
60	test	19	1,1-dichloroethane	75-34-3	C(Cl)(Cl)C	1.490	1.72	1.84	1.25	1.82	1.76	1.82	1.94
61	training	42	1,2-dichloroethane	107-06-2	ClCCCl	1.650	1.48	1.74	1.51	1.82	1.83	1.74	1.48
63	test	20	dichloroethane	1300-21-6	ClCCCl	1.785	1.48	1.74	1.51	1.82	1.83	1.74	1.48
64	training	43	1,2-diiodoethane	624-73-7	C(CI)I	2.851	2.72	2.52	2.86	2.60	2.84	2.82	2.71
65	training	44	acetaldehyde	75-07-0	O=CC	1.622	-0.01	0.43	-0.18	-0.32	-0.17	0.33	-0.27
66	training	45	ethylene oxide	75-21-8	O(C1)C1	1.214	-0.47	0.25	-0.13	-0.56	-0.05	-0.19	-0.15
67	test	21	acetic acid	64-19-7	O=C(O)C	1.285	-0.12	0.03	-0.23	-0.39	0.09	-0.08	-0.21
68	test	22	methyl formate	107-31-3	O=COC	1.393	-0.31	0.20	-0.03	-0.39	-0.17	0.00	0.03
69	training	46	bromoethane	74-96-4	BrCC	2.247	1.64	1.61	1.34	1.60	1.67	1.74	1.35
70	test	23	ethyl chloride	75-00-3	ClCC	2.155	1.47	1.56	1.19	1.36	1.58	1.40	1.18
71	training	47	2-chloroethanol	107-07-3	OCCCl	1.393	0.00	0.61	0.31	0.35	0.11	0.27	-0.06
72	training	48	ethyl iodide	75-03-6	C(C)I	2.465	2.29	1.95	1.87	1.82	2.08	1.94	2.05
73	training	49	acetamide	60-35-5	O=C(N)C	0.692	-1.10	-0.51	-0.83	-0.79	-1.16	-0.81	-0.86
74	training	50	N-methylformamide	123-39-7	O=CNC	0.849	-1.31	-0.26	-0.67	-0.79	-1.14	-0.53	-0.97

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
75	test	24	nitroethane	79-24-3	N(=O)(=O)CC	1.475	0.45	0.40	0.64	0.21	0.45	0.60	0.18
76	training	51	ethane	74-84-0	CC	2.362	1.44	1.38	1.28	1.76	1.32	1.48	1.30
77	test	25	ethyl alcohol	64-17-5	OCC	1.214	-0.40	0.43	-0.01	-0.17	-0.14	-0.08	-0.09
78	training	52	dimethyl ether	115-10-6	O(C)C	1.431	-0.16	0.44	0.05	-0.17	0.07	0.02	0.08
79	training	53	dimethyl sulfoxide	67-68-5	O=S(C)C	0.643	-1.09	-0.91	-0.32	-0.32	-1.22	-0.72	-0.61
80	training	54	ethylene glycol	107-21-1	OCCO	0.637	-1.53	-0.53	-0.90	-1.05	-1.20	-1.21	-1.36
81	training	55	dimethyl sulfone	67-71-0	O=S(=O)(C)C	0.610	-0.95	-0.61	-0.22	-0.49	-1.11	0.09	-0.40
82	training	56	dimethyl sulfate	77-78-1	O=S(=O)(OC)OC	2.008	-0.60	-0.61	-0.53	-0.51	0.16	-0.86	-0.30
83	training	57	dimethyl disulfide	624-92-0	S(SC)C	2.340	1.15	-0.77	1.39	0.85	1.87	1.57	1.77
84	test	26	ethylamine	75-04-7	NCC	1.306	-0.20	-0.11	-0.30	-0.17	-0.15	-0.14	-0.35
85	test	27	dimethylamine	124-40-3	N(C)C	2.720	-0.53	-0.03	-0.22	-0.17	-0.17	-0.18	-0.20
86	test	28	monoethanolamine	141-43-5	OCCN	0.664	-1.53	-1.07	-1.19	-1.05	-1.61	-1.27	-1.31
87	training	58	ethylenediamine	107-15-3	NCCN	0.267	-1.77	-1.61	-1.48	-1.05	-1.62	-1.34	-2.04
88	training	59	cyanogen	460-19-5	C(#N)C(#N)	1.415	-0.65	0.48	0.07	-1.31	0.07	-0.47	0.07
89	training	60	hexafluoroacetone	684-16-2	O=C(C(F)(F)F)C(F)(F)F	2.171	1.77	1.71	1.93	1.44	0.60	0.76	1.46
90	training	61	malononitrile	109-77-3	C(#N)CC(#N)	1.051	-0.84	0.94	0.10	-0.79	-0.60	-0.08	-0.50
91	test	29	acrylonitrile	107-13-1	C(#N)C=C	1.513	0.20	1.10	0.74	0.09	0.21	0.32	0.25
92	training	62	oxazole	288-42-6	c1cocn1	1.442	-0.09	0.35	-0.27	-0.37	0.21	-0.15	0.12
93	training	63	thiazole	288-47-1	s1cncc1	1.616	0.89	0.62	0.29	-0.30	0.99	0.60	0.44
94	test	30	methylacetylene	74-99-7	C(#C)C	1.888	0.92	0.83	1.92	1.22	1.04	1.02	0.91
95	training	64	allene	463-49-0	C=C=C	2.166	1.67	1.15	1.24	1.11	1.65	3.34	0.81
96	training	65	cis-1,2-dichloropropene	6923-20-2	CC(=CCl)Cl	2.481	2.10	2.63	1.57	2.08	2.53	2.11	2.13
97	training	66	imidazole	288-32-4	N1C=NC=C1	1.333	-0.21	0.02	-0.28	-0.37	0.06	-0.04	-0.08
98	training	67	1H-pyrazole	288-13-1	N1N=CC=C1	1.448	0.03	-0.07	0.27	-0.37	0.06	0.79	0.26
99	training	68	acrolein	107-02-8	O=CC=C	1.372	0.18	0.60	0.51	0.09	0.19	0.29	-0.01
100	test	31	propargyl alcohol	107-19-7	OCC#C	1.170	-0.70	-0.12	0.83	0.20	-0.42	-0.23	-0.38
101	test	32	acrylic acid	79-10-7	O=C(O)C=C	1.567	0.46	0.19	0.47	0.03	0.44	0.21	0.35
102	test	33	3-bromo-1-propene	106-95-6	BrCC=C	2.351	1.98	1.78	1.61	1.88	2.02	1.73	1.79
103	test	34	2-chloro-1-propene	557-98-2	C(=C)(C)Cl	2.465	1.88	2.22	1.52	1.67	2.17	1.40	1.86
104	training	69	α-epichlorohydrin	106-89-8	O(C1CCl)C1	1.540	0.35	0.71	0.56	0.41	0.63	0.61	0.45
105	training	70	1,2,3-trichloropropane	96-18-4	ClCC(Cl)CCl	2.612	2.29	2.01	2.20	2.60	2.50	2.38	1.79
106	test	35	propionitrile	107-12-0	C(#N)CC	1.464	-0.01	1.39	0.71	0.20	0.35	0.28	0.16
107	training	71	acrylamide	79-06-1	O=C(N)C=C	0.953	-0.65	-0.34	-0.13	-0.38	-0.81	-0.52	-0.67
108	test	36	lactonitrile	78-97-7	N#CC(O)C	0.866	-0.65	0.38	-0.04	-0.68	-1.19	-0.48	-0.27
109	training	72	nitroglycerine	55-63-0	O(N(=O)=O)CC(O(N(=O)=O))CO(N(=O)=O	2.258	1.25	2.01	0.08	0.57	1.51	0.98	1.62
110	test	37	cyclopropane	75-19-4	C(C1)C1	2.313	1.56	1.45	1.37	1.88	1.70	1.71	1.62
111	test	38	propylene	115-07-1	C(=C)C	2.340	1.68	1.55	1.35	1.22	1.68	1.79	1.44
112	training	73	1,2-dichloropropane	78-87-5	ClCC(Cl)C	2.465	2.13	1.83	1.89	2.23	2.25	2.04	1.84
113	training	74	allyl alcohol	107-18-6	OCC=C	1.469	-0.03	0.60	0.26	0.20	0.21	0.12	0.17
114	training	75	propionaldehyde	123-38-6	O=CCC	1.698	0.31	0.89	0.48	0.20	0.33	0.59	0.59
115	test	39	acetone	67-64-1	O=C(C)C	1.246	-0.29	0.74	-0.24	0.20	-0.24	0.19	-0.05
116	training	76	1,2-propylene oxide	75-56-9	O(C1C)C1	1.393	0.04	0.53	0.25	-0.05	0.37	0.27	0.03

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
117	training	77	1,3-propylene oxide	503-30-0	O(CC1)C1	1.301	0.05	0.56	-0.07	-0.05	0.45	0.17	-0.14
119	training	78	propanoic acid	79-09-4	O=C(O)CC	1.557	0.31	0.49	0.44	0.13	0.58	0.17	0.33
120	training	79	ethyl formate	109-94-4	O=COCC	1.502	0.38	0.64	0.32	0.13	0.32	0.43	0.50
121	training	80	methyl acetate	79-20-9	O=C(OC)C	1.475	0.18	0.48	0.02	0.13	0.37	0.24	0.18
122	test	40	3-mercaptopropionic acid	107-96-0	O=C(O)CCS	1.611	0.34	0.62	0.32	0.13	0.52	0.42	0.43
123	test	41	lactic acid	50-21-5	O=C(O)C(O)C	0.985	-0.79	-0.52	-0.31	-0.70	-0.65	-0.59	-0.72
124	test	42	trioxane	110-88-3	O(COCO1)C1	1.143	-0.95	-0.22	-0.45	-0.54	-0.56	-0.16	-0.43
125	training	81	1-bromopropane	106-94-5	BrCCC	2.519	2.18	2.07	1.86	2.03	2.16	2.09	2.10
126	test	43	2-bromopropane	75-26-3	BrC(C)C	2.411	1.83	1.70	1.72	2.03	2.08	2.03	1.79
127	test	44	1-chloropropane	540-54-5	ClCCC	2.487	2.09	2.02	1.72	1.82	2.07	1.76	2.04
128	training	82	2-chloropropane	75-29-6	C(Cl)(C)C	2.411	1.49	1.65	1.57	1.82	2.00	1.70	1.62
129	training	83	1-iodopropane	107-08-4	C(CC)I	2.737	2.65	2.41	2.39	2.23	2.57	2.29	2.59
130	training	84	2-iodopropane	75-30-9	C(C)(C)I	2.949	2.59	2.04	2.25	2.23	2.50	2.23	2.30
131	test	45	allylamine	107-11-9	NCC=C	1.393	-0.43	0.06	-0.03	0.20	0.21	0.05	0.07
132	training	85	N,N-dimethylformamide	68-12-2	O=CN(C)C	0.828	-0.77	-0.05	-0.47	-0.27	-0.93	-0.38	-1.01
133	training	86	N-methylacetamide	79-16-3	O=C(NC)C	0.806	-1.06	0.01	-0.63	-0.27	-0.70	-0.29	-1.05
134	training	87	1-nitropropane	108-03-2	N(=O)(=O)CCC	1.850	0.91	0.86	1.16	0.73	0.95	0.96	0.87
135	training	88	2-nitropropane	79-46-9	N(=O)(=O)C(C)C	1.883	0.71	0.80	1.02	0.73	0.87	1.13	0.80
136	training	89	propane	74-98-6	C(C)C	2.661	2.19	1.84	1.74	2.28	1.81	2.05	1.84
137	training	90	glyphosate	1071-83-6	OC(=O)CNCP(O)(O)=O	3.460	-2.43	-4.81	-2.07	-1.96	-4.47	-2.68	-4.62
138	training	91	propyl alcohol	71-23-8	OCCC	1.513	0.21	0.89	0.51	0.35	0.35	0.28	0.25
139	training	92	isopropyl alcohol	67-63-0	OC(C)C	1.404	0.04	0.83	0.37	0.35	0.28	0.38	0.34
140	training	93	2-methoxyethanol	109-86-4	O(CCO)C	0.958	-0.78	-0.07	-0.49	-0.53	-0.91	-0.69	-0.77
141	training	94	1,2-propanediol	57-55-6	OCC(O)C	0.877	-1.10	-0.12	-0.52	-0.53	-0.78	-0.75	-0.92
142	test	46	1,3-propanediol	504-63-2	OCCCO	0.811	-1.18	-0.06	-0.83	-0.53	-0.71	-0.85	-1.04
143	test	47	glycerol	56-81-5	OCC(O)CO	0.420	-1.93	-1.08	-1.41	-1.37	-1.65	-1.88	-1.76
144	test	48	propyl mercaptan	107-03-9	SCCC	2.362	1.72	1.97	1.49	1.36	1.76	1.66	1.81
145	training	95	methyl ethyl sulfide	624-89-5	S(CC)C	2.215	1.16	1.55	1.14	1.36	1.41	1.67	1.54
146	training	96	propylamine	107-10-8	NCCC	1.638	0.31	0.35	0.22	0.35	0.34	0.22	0.48
147	training	97	isopropylamine	75-31-0	NC(C)C	1.518	-0.05	0.29	0.08	0.35	0.27	0.32	0.09
148	training	98	methylethylamine	624-78-2	N(CC)C	1.459	0.13	0.41	0.13	0.35	0.32	0.25	0.15
149	training	99	trimethylamine	75-50-3	N(C)(C)C	1.464	-0.14	0.19	0.32	0.35	0.04	0.06	0.26
150	training	100	1-amino-2-propanol	78-96-6	OC(CN)C	0.855	-1.03	-0.66	-0.81	-0.53	-1.19	-0.81	-0.96
151	test	49	3-amino-1-propanol	156-87-6	OCCCN	0.768	-1.01	-0.60	-1.13	-0.53	-1.12	-0.92	-1.12
152	training	101	methylethanolamine	109-83-1	OCCNC	0.866	-1.05	-0.54	-0.75	-0.53	-1.15	-0.88	-0.94
153	training	102	trimethyl phosphate	512-56-1	O=P(OC)(OC)OC	1.023	-0.61	-0.50	-0.35	0.11	-0.60	-0.69	-0.48
154	training	103	hexachloro-1,3-butadiene	87-68-3	C(=C(C(=C(Cl)Cl)Cl)Cl)(Cl)Cl	3.977	4.86	6.17	3.63	3.66	4.72	4.38	4.78
155	test	50	succinonitrile	110-61-2	C(#N)CCC(#N)	0.838	-0.75	1.41	0.14	-0.33	-0.63	-0.49	-0.99
156	training	104	pyrimidine	289-95-2	n(cccn1)c1	1.159	-0.21	0.27	0.05	-0.01	-0.06	-0.27	-0.40
157	test	51	furan	110-00-9	O1C=CC=C1	2.106	1.24	0.98	0.93	0.16	1.36	0.71	1.34
158	test	52	fumaric acid	110-17-8	O=C(O)C=CC(=O)O	1.627	0.21	-0.76	-0.01	-0.45	0.05	-0.42	-0.34
159	training	105	maleic acid	110-16-7	O=C(O)C=CC(=O)O	1.116	0.21	-0.76	-0.01	-0.45	0.05	-0.42	-0.34

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
160	training	106	thiophene	110-02-1	S1C=CC=C1	2.362	1.89	1.69	1.49	1.17	1.81	1.46	1.81
161	training	107	methacrylonitrile	126-98-7	N#CC(=C)C	1.747	0.91	1.57	1.19	0.55	0.76	0.46	0.68
162	test	53	vinylacetonitrile	109-75-1	N#CCC=C	1.595	0.61	1.56	0.78	0.55	0.70	0.68	0.40
163	training	108	pyrrole	109-97-7	C1=CC=CN1	1.785	0.76	0.66	0.92	0.16	0.88	0.82	0.75
164	training	109	methyl cyanoacetate	105-34-0	O=C(OC)CC#N	1.121	-0.10	0.49	0.08	-0.36	-0.47	0.13	-0.47
165	training	110	dimethylacetylene	503-17-3	C(#CC)C	2.171	1.70	1.86	1.87	1.67	1.59	1.62	1.46
166	training	111	1,3 butadiene	106-99-0	C(C=C)=C	2.460	1.94	1.72	1.41	1.57	2.03	1.96	1.99
167	training	112	2,5-dihydrofuran	1708-29-8	O(CC=C1)C1	1.627	0.48	0.78	0.34	0.26	0.72	0.41	0.46
168	training	113	γ-butyrolactone	96-48-0	O=C(OCC1)C1	1.029	-0.11	0.27	0.28	0.19	-0.31	0.22	-0.64
169	training	114	methacrylic acid	79-41-4	O=C(O)C(=C)C	1.883	0.63	0.67	0.91	0.48	0.99	0.36	0.93
170	test	54	methyl acrylate	96-33-3	O=C(OC)C=C	1.812	0.67	0.65	0.72	0.48	0.73	0.53	0.80
171	training	115	vinyl acetate	108-05-4	O=C(OC=C)C	1.774	0.83	1.09	0.10	0.48	0.73	0.69	0.73
172	training	116	succinic acid	110-15-6	O=C(O)CCC(=O)O	1.056	-0.53	-0.40	-0.41	-0.35	-0.75	-0.71	-0.59
173	training	117	butyronitrile	109-74-0	C(#N)CCC	1.703	0.59	1.86	1.17	0.65	0.84	0.85	0.53
174	training	118	isobutyronitrile	78-82-0	N#CC(C)C	1.627	0.50	1.73	1.17	0.65	0.76	0.58	0.46
175	training	119	2-pyrrolidone	616-45-5	O=C(NCC1)C1	0.915	-0.90	0.10	-0.37	-0.21	-0.32	-0.31	-0.82
176	training	120	1-butene	106-98-9	C(=C)CC	2.683	2.21	2.01	1.81	1.67	2.17	2.25	2.40
177	training	121	cis-2-butene	590-18-1	C(=CC)C	2.645	2.32	1.95	1.75	1.67	2.09	2.31	2.33
178	training	122	trans-2-butene	624-64-6	C(=CC)C	2.634	2.32	1.95	1.75	1.67	2.09	2.31	2.33
179	training	123	isobutene	115-11-7	C(=C)(C)C	2.655	1.87	2.02	1.80	1.67	2.23	1.85	2.06
180	training	124	bis(2-chloroethyl) ether	111-44-4	O(CCCl)CCCl	1.986	1.23	1.67	1.38	1.59	1.56	1.55	1.29
181	training	125	ethyl vinyl ether	109-92-2	O(C=C)CC	1.943	1.19	1.49	0.48	0.65	0.91	1.25	1.04
182	training	126	butyraldehyde	123-72-8	O=CCCC	1.856	1.10	1.36	0.94	0.65	0.82	1.16	0.88
183	training	127	methyl ethyl ketone	78-93-3	O=C(CC)C	1.535	0.41	1.21	0.42	0.65	0.26	0.44	0.29
184	test	55	tetrahydrofuran	109-99-9	O(CCC1)C1	1.627	0.35	0.88	0.51	0.41	0.94	0.53	0.46
185	training	128	butyric acid	107-92-6	O=C(O)CCC	1.807	0.78	0.95	0.89	0.59	1.07	0.74	0.79
186	training	129	isobutyric acid	79-31-2	O=C(O)C(C)C	1.888	0.78	0.83	0.90	0.59	1.00	0.47	0.83
187	training	130	propyl formate	110-74-7	O=COCCC	1.829	0.93	1.11	0.85	0.59	0.81	0.79	0.83
188	test	56	ethyl acetate	141-78-6	O=C(OCC)C	1.774	0.74	0.91	0.37	0.59	0.86	0.66	0.73
189	training	131	methyl propanoate	554-12-1	O=C(OC)CC	1.834	0.68	0.94	0.69	0.59	0.86	0.49	0.82
190	training	132	1,4-dioxane	123-91-1	O(CCOC1)C1	1.149	-0.23	-0.01	-0.26	-0.47	-0.32	-0.38	-0.27
191	test	57	sulfolane	126-33-0	O=S(=O)(CCC1)C1	0.958	-0.65	0.33	0.24	0.09	-0.24	0.47	-0.77
192	test	58	1-bromobutane	109-65-9	BrCCCC	2.873	2.73	2.54	2.32	2.42	2.65	2.66	2.75
193	training	133	1-chlorobutane	109-69-3	ClCCCC	2.813	2.37	2.49	2.17	2.23	2.56	2.33	2.64
194	test	59	2-chlorobutane	78-86-4	C(Cl)(CC)C	2.645	2.34	2.12	2.10	2.23	2.49	2.05	2.33
195	training	134	1-fluorobutane	2366-52-1	FCCCC	2.781	1.79	2.08	1.91	2.03	2.25	2.19	2.58
196	test	60	1-iodobutane	542-69-8	C(CCC)I	3.009	3.11	2.88	2.85	2.60	3.06	2.86	3.08
197	training	135	pyrrolidine	123-75-1	N(CCC1)C1	1.627	0.16	0.72	0.25	0.41	0.70	0.33	0.46
198	training	136	N,N-dimethylacetamide	127-19-5	O=C(N(C)C)C	0.958	-0.59	0.23	-0.42	0.18	-0.49	-0.15	-0.77
199	training	137	morpholine	110-91-8	O(CCNC1)C1	0.398	-0.75	-0.17	-0.53	-0.47	-0.56	-0.58	-0.86
200	training	138	butanamide	541-35-5	O=C(N)CCC	1.263	-0.13	0.41	0.29	0.18	-0.18	0.01	-0.21
201	training	139	1-nitrobutane	627-05-4	CCCCN(=O)=O	2.177	1.49	1.32	1.62	1.18	1.44	1.53	1.47

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
202	training	140	butane	106-97-8	C(CC)C	2.949	2.81	2.31	2.20	2.73	2.31	2.62	2.89
203	training	141	piperazine	110-85-0	N(CCNC1)C1	0.741	-1.16	-0.34	-0.79	-0.47	-0.80	-0.77	-1.50
204	training	142	butanol	71-36-3	OCCCC	1.834	0.84	1.35	0.97	0.80	0.84	0.85	0.88
205	test	61	isobutanol	78-83-1	OCC(C)C	1.790	0.60	1.23	0.83	0.80	0.77	0.58	0.76
206	training	143	sec-butanol	78-92-2	OC(CC)C	1.731	0.66	1.29	0.89	0.80	0.77	0.74	0.61
207	training	144	tert-butanol	75-65-0	OC(C)(C)C	1.567	0.70	0.98	0.57	0.80	0.73	0.80	0.53
208	test	62	diethyl ether	60-29-7	O(CC)CC	1.861	1.12	1.31	0.75	0.80	1.05	0.87	0.89
209	training	145	methyl propyl ether	557-17-5	CCCOC	2.035	0.90	1.34	0.92	0.80	1.05	0.80	1.21
210	test	63	1,2-dimethoxyethane	110-71-4	O(CCOC)C	1.263	0.03	0.38	-0.08	-0.08	-0.21	-0.17	-0.21
211	test	64	2-ethoxyethanol	110-80-5	O(CCO)CC	1.203	-0.28	0.36	-0.14	-0.08	-0.42	-0.27	-0.32
212	training	146	1,4-butanediol	110-63-4	OCCCCO	0.925	-0.63	0.40	-0.25	-0.08	-0.22	-0.49	-0.83
213	training	147	diethyl sulfate	64-67-5	O=S(=O)(OCC)OCC	1.997	-0.29	0.26	0.17	0.46	1.14	-0.02	1.14
214	test	65	butyl mercaptan	109-79-5	SCCCC	2.617	2.51	2.43	1.94	1.82	2.25	2.23	2.28
215	test	66	diethyl sulfide	352-93-2	S(CC)CC	2.438	2.46	2.06	1.49	1.82	1.90	2.03	1.95
216	training	148	butylamine	109-73-9	NCCCC	1.845	0.85	0.81	0.68	0.80	0.83	0.78	0.97
217	training	149	isobutylamine	78-81-9	NCC(C)C	1.774	0.54	0.69	0.54	0.80	0.76	0.51	0.73
218	training	150	tert-butylamine	75-64-9	NC(C)(C)C	1.595	0.81	0.44	0.28	0.80	0.72	0.73	0.27
219	test	67	diethylamine	109-89-7	N(CC)CC	1.693	0.76	0.84	0.48	0.80	0.81	0.67	0.58
220	test	68	diethanolamine	111-42-2	OCCNCCO	0.599	-1.41	-1.06	-1.29	-0.92	-1.71	-1.59	-1.43
221	training	151	hexachlorocyclopentadiene	77-47-4	C(=C(C(=C1Cl)Cl)Cl)(C1(Cl)Cl)Cl	4.119	4.85	4.39	3.50	3.56	4.63	3.62	5.04
222	training	152	furfural	98-01-1	O=CC(OC=C1)=C1	1.600	0.43	0.77	0.99	-0.14	0.83	0.42	0.41
223	test	69	pyridine	110-86-1	n(cccc1)c1	1.731	0.70	0.90	0.68	0.47	0.80	0.77	0.65
224	training	153	glutaronitrile	544-13-8	N#CCCCC#N	0.985	-0.49	1.87	0.60	0.08	-0.14	-0.13	-0.72
225	training	154	2-methylfuran	534-22-5	C1=C(C)OC=C1	2.383	1.75	1.40	1.08	0.56	1.91	0.94	1.85
226	test	70	furfuryl alcohol	98-00-0	C1=C(CO)OC=C1	1.529	0.25	0.46	0.62	-0.32	0.45	-0.09	0.28
227	training	155	2-methylthiophene	554-14-3	C1=C(C)SC=C1	2.645	2.30	2.14	1.64	1.58	2.36	1.31	2.33
228	test	71	3-methylthiophene	616-44-4	C1(C)=CSC=C1	2.650	2.28	2.01	1.98	1.58	2.36	1.60	2.34
229	test	72	N-methylpyrrole	96-54-8	C1=CN(C)C=C1	2.035	1.31	0.51	1.13	0.56	1.43	1.12	1.21
230	training	156	isoprene	78-79-5	C(C=C)(=C)C	2.693	2.22	2.19	1.86	1.97	2.58	2.31	2.47
231	training	157	cis-1,3-pentadiene	1574-41-0	C(=CC=C)C	2.683	2.65	2.12	1.82	1.97	2.45	2.48	2.40
232	training	158	trans-1,3-pentadiene	2004-70-8	C(=CC=C)C	2.704	2.65	2.12	1.82	1.97	2.45	2.48	2.40
233	training	159	1,4-pentadiene	591-93-5	C(=C)CC=C	2.726	2.39	2.18	1.87	1.97	2.52	2.64	2.48
234	training	160	1-pentyne	627-19-0	C(#C)CCC	2.454	2.13	1.76	2.84	2.08	2.03	2.05	1.98
235	training	161	acetylacetone	123-54-6	O=C(CC(=O)C)C	1.595	-0.20	0.57	-0.47	0.08	0.05	0.23	0.40
236	test	73	allyl acetate	591-87-7	O=C(OCC=C)C	1.905	1.03	1.08	0.64	0.89	1.22	0.86	0.97
237	training	162	ethyl acrylate	140-88-5	O=C(OCC)C=C	2.095	1.24	1.08	1.07	0.89	1.22	0.96	1.32
238	training	163	methyl methacrylate	80-62-6	O=C(OC)C(=C)C	2.128	1.10	1.12	1.16	0.89	1.28	0.68	1.38
239	training	164	2-hydroxyethyl acrylate	818-61-1	O=C(OCCO)C=C	1.263	0.04	0.13	0.18	0.05	-0.25	-0.18	-0.21
240	test	74	levulinic acid	123-76-2	O=C(O)CCC(=O)C	1.110	-0.14	0.32	-0.42	0.05	-0.49	-0.44	-0.49
241	training	165	glutaric acid	110-94-1	O=C(O)CCCC(=O)O	1.219	-0.25	0.06	0.05	0.06	-0.26	-0.35	-0.29
242	training	166	valeronitrile	110-59-8	N#CCCCC	1.888	1.10	2.32	1.62	1.06	1.33	1.42	1.12
243	training	167	N-methyl-2-pyrrolidone	872-50-4	O=C1CCCN1C	1.170	-0.72	0.39	-0.17	0.20	-0.11	-0.17	-0.54

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
244	test	75	L-glutamic acid	56-86-0	O=C(O)C(N)CCC(=O)O	-0.630	-3.54	-1.49	-0.92	-2.95	-3.83	-3.35	-3.69
245	training	168	cyclopentane	287-92-3	C(CCC1)C1	3.009	2.88	2.09	2.28	2.75	2.68	2.85	3.00
246	test	76	methyl propyl ketone	107-87-9	O=C(CCC)C	1.834	0.87	1.67	0.88	1.06	0.75	1.01	0.91
247	training	169	diethyl ketone	96-22-0	O=C(CC)CC	1.823	1.19	1.67	1.09	1.06	0.75	0.69	0.89
248	training	170	methyl isopropyl ketone	563-80-4	O=C(C(C)C)C	1.682	0.78	1.55	0.88	1.06	0.67	0.73	0.84
249	test	77	2-methyltetrahydrofuran	96-47-9	O(C(CC1)C)C1	2.383	0.96	1.16	0.89	0.82	1.35	0.99	1.00
250	test	78	tetrahydropyran	142-68-7	O(CCCC1)C1	1.823	1.16	1.20	0.97	0.82	1.43	0.88	0.95
251	training	171	pentanoic acid	109-52-4	O=C(O)CCCC	2.133	1.34	1.42	1.35	1.00	1.56	1.31	1.39
252	training	172	3-methylbutanoic acid	503-74-2	O=C(O)CC(C)C	2.008	1.26	1.29	1.14	1.00	1.49	1.25	1.16
253	test	79	propyl acetate	109-60-4	O=C(OCCC)C	2.052	1.28	1.38	0.89	1.00	1.36	1.02	1.24
254	training	173	ethyl propanoate	105-37-3	O=C(OCC)CC	2.035	1.32	1.38	1.04	1.00	1.36	0.92	1.21
255	training	174	methyl butanoate	623-42-7	O=C(OC)CCC	2.079	1.22	1.41	1.14	1.00	1.36	1.06	1.29
256	training	175	diethyl carbonate	105-58-8	O=C(OCC)OCC	2.035	0.86	1.61	1.36	0.57	1.22	1.26	1.21
257	training	176	1-bromopentane	110-53-2	BrCCCCC	3.210	3.27	3.00	2.78	2.78	3.14	3.23	3.37
258	test	80	1-chloropentane	543-59-9	ClCCCCC	2.862	3.12	2.95	2.63	2.60	3.05	2.89	2.61
259	training	177	2-chloro-2-methylbutane	594-36-5	C(CC)(Cl)(C)C	2.748	2.95	3.08	2.30	2.60	2.94	2.45	2.52
260	training	178	1-fluoropentane	592-50-7	FCCCCC	2.645	2.93	2.54	2.37	2.42	2.74	2.76	2.33
261	test	81	N-methylpyrrolidine	120-94-5	N(CCC1)(C1)C	1.877	0.54	1.00	0.78	0.82	0.91	0.57	0.92
262	training	179	piperidine	110-89-4	N(CCCC1)C1	1.834	0.97	1.03	0.70	0.82	1.19	0.69	0.84
263	training	180	1-nitropentane	628-05-7	CCCCCN(=O)(=O)	2.470	2.00	1.79	2.08	1.59	1.93	2.10	2.01
264	training	181	pentane	109-66-0	C(CCC)C	3.254	3.41	2.77	2.65	3.14	2.80	3.19	3.39
265	training	182	isopentane	78-78-4	C(CC)(C)C	2.628	3.12	2.65	2.45	3.14	2.72	3.12	2.64
266	training	183	neopentane	463-82-1	C(C)(C)(C)C	3.069	2.95	2.70	2.20	3.14	2.69	3.16	2.49
267	training	184	dimethoate	60-51-5	O=C(NC)CSP(OC)(OC)=S	2.560	1.21	0.66	0.57	-0.75	0.28	0.90	0.78
268	training	185	1-pentanol	71-41-0	OCCCCC	2.198	1.47	1.82	1.43	1.21	1.33	1.42	1.56
269	training	186	2-pentanol	6032-29-7	OC(CCC)C	2.057	1.18	1.76	1.35	1.21	1.26	1.31	1.19
270	test	82	3-pentanol	584-02-1	OC(CC)CC	2.035	1.22	1.76	1.42	1.21	1.26	1.10	1.21
271	training	187	2-methyl-1-butanol	137-32-6	OCC(CC)C	2.079	1.24	1.69	1.29	1.21	1.26	1.15	1.23
272	training	188	3-methyl-1-butanol	123-51-3	OCCC(C)C	2.073	1.33	1.69	1.22	1.21	1.26	1.36	1.16
273	training	189	tert-pentyl-alcohol	75-85-4	OC(CC)(C)C	1.861	1.19	1.44	1.10	1.21	1.22	1.16	0.89
274	test	83	3-methyl-2-butanol	598-75-4	OC(C(C)C)C	2.073	0.89	1.63	1.21	1.21	1.19	1.04	1.28
275	training	190	2,2-dimethyl-1-propanol	75-84-3	OCC(C)(C)C	2.090	1.15	1.75	1.11	1.21	1.22	0.97	1.31
276	test	84	methyl tert-butyl ether	1634-04-4	O(C(C)(C)C)C	1.888	1.53	1.43	0.98	1.21	1.43	1.32	0.94
277	training	191	pentaerythritol	115-77-5	OCC(CO)(CO)CO	0.458	-1.92	-1.11	-2.16	-1.32	-1.77	-2.42	-2.39
278	training	192	pentylamine	110-58-7	NCCCCC	2.188	1.39	1.28	1.14	1.21	1.33	1.35	1.49
279	training	193	hexachlorobenzene	118-74-1	c(c(c(c(c1Cl)Cl)Cl)Cl)(c1Cl)Cl	4.490	5.70	5.66	5.82	5.21	5.86	5.75	5.73
280	training	194	hexafluorobenzene	392-56-3	Fc(c(F)c(F)c(F)c1F)c1F	2.764	2.33	2.34	3.06	4.33	3.20	2.99	2.55
281	test	85	pentachlorobenzene	608-93-5	c(c(c(c(c1Cl)Cl)Cl)Cl)(c1)Cl	4.113	5.22	5.05	5.15	4.93	5.22	5.13	5.18
282	test	86	pentachlorophenol	87-86-5	Oc(c(c(c(c1Cl)Cl)Cl)Cl)c1Cl	2.470	4.99	4.75	4.89	3.91	4.74	4.73	5.12
283	training	195	1,2,3,4-tetrachlorobenzene	634-66-2	c(c(c(c(c1)Cl)Cl)Cl)(c1)Cl	3.520	4.62	4.43	4.49	4.63	4.57	4.51	4.64
284	training	196	1,2,3,5-tetrachlorobenzene	634-90-2	c(cc(c(c1Cl)Cl)Cl)(c1)Cl	3.520	4.63	4.43	4.49	4.63	4.57	4.51	4.66
285	test	87	1,2,4,5-tetrachlorobenzene	95-94-3	c(c(cc(c1Cl)Cl)Cl)(c1)Cl	3.720	4.61	4.43	4.49	4.63	4.57	4.51	4.60

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
286	training	197	2,3,4,5-tetrachlorophenol	4901-51-3	Oc1cc(Cl)c(Cl)c(Cl)c1Cl	2.880	4.41	4.14	4.22	3.62	4.09	4.10	4.21
287	test	88	2,3,4,6-tetrachlorophenol	58-90-2	Oc(c(cc(c1Cl)Cl)Cl)c1Cl	2.880	4.37	4.14	4.22	3.62	4.09	4.10	4.45
288	training	198	2,3,5,6-tetrachlorophenol	935-95-5	Oc1c(Cl)c(Cl)cc(Cl)c1Cl	2.880	4.45	4.14	4.22	3.62	4.09	4.10	3.88
289	test	89	1-chloro-2,4-dinitrobenzene	97-00-7	N(=O)(=O)c(ccc(c1N(=O)(=O))Cl)c1	2.557	2.29	2.61	2.28	2.51	2.27	2.43	2.29
290	test	90	1,2-dichloro-4-nitrobenzene	99-54-7	N(=O)(=O)c(ccc(c1Cl)Cl)c1	2.530	3.11	3.21	3.05	3.11	3.10	3.16	3.12
291	test	91	1,2,4-trichlorobenzene	120-82-1	c(ccc(c1Cl)Cl)(c1)Cl	3.110	4.08	3.82	3.82	4.06	3.93	3.89	4.02
292	test	92	1,2,3-trichlorobenzene	87-61-6	c(c(c(cc1)Cl)Cl)(c1)Cl	3.230	4.07	3.82	3.82	4.06	3.93	3.89	4.14
293	training	199	1,3,5-trichlorobenzene	108-70-3	c(cc(cc1Cl)Cl)(c1)Cl	2.850	4.08	3.82	3.82	4.06	3.93	3.89	4.19
295	training	200	2,3,4-trichlorophenol	15950-66-0	Oc1ccc(Cl)c(Cl)c1Cl	1.960	3.78	3.52	3.56	3.31	3.45	3.48	3.47
296	training	201	2,3,5-trichlorophenol	933-78-8	Oc1cc(Cl)cc(Cl)c1Cl	1.960	3.77	3.52	3.56	3.31	3.45	3.48	3.58
297	test	93	2,3,6-trichlorophenol	933-75-5	Oc(c(ccc1Cl)Cl)c1Cl	1.960	3.77	3.52	3.56	3.31	3.45	3.48	3.77
298	training	202	2,4,5-trichlorophenol	95-95-4	Oc(c(cc(c1Cl)Cl)Cl)c1	1.960	3.79	3.52	3.56	3.31	3.45	3.48	3.72
299	training	203	2,4,6-trichlorophenol	88-06-2	Oc(c(cc(c1)Cl)Cl)c1Cl	1.960	3.78	3.52	3.56	3.31	3.45	3.48	3.69
300	test	94	3,4,5-trichlorophenol	609-19-8	Oc1cc(Cl)c(Cl)c(Cl)c1	1.960	3.77	3.52	3.56	3.31	3.45	3.48	4.01
301	test	95	nitrapyrin	1929-82-4	n(c(ccc1)C(Cl)(Cl)Cl)c1Cl	2.240	3.87	3.42	3.51	2.69	3.35	2.91	3.41
302	training	204	1,3,5-trinitrobenzene	99-35-4	O=N(=O)c(cc(N(=O)=O)cc1N(=O)=O)c1	2.019	1.54	1.59	1.51	2.01	1.45	1.70	1.18
303	test	96	1-bromo-2-chlorobenzene	694-80-4	c(c(ccc1)Br)(c1)Cl	2.600	3.61	3.29	3.24	3.64	3.53	3.44	3.24
304	training	205	1-bromo-3-chlorobenzene	108-37-2	c(cccc1Br)(c1)Cl	2.600	3.59	3.29	3.24	3.64	3.53	3.44	3.59
305	training	206	1-bromo-4-chlorobenzene	106-39-8	c(ccc(c1)Br)(c1)Cl	2.600	3.63	3.29	3.24	3.64	3.53	3.44	3.50
307	test	97	3-bromo-5-chlorophenol	56962-04-0	Oc1cc(Cl)cc(Br)c1	2.600	3.14	2.99	2.97	2.89	3.05	3.04	3.68
308	training	207	4-bromo-2-chlorophenol	3964-56-5	Oc1c(Cl)cc(Br)cc1	2.600	3.15	2.99	2.97	2.89	3.05	3.04	2.45
309	training	208	2-bromo-4-chlorophenol	695-96-5	Brc(cc(Cl)c1)c(c1)O	2.600	3.12	2.99	2.97	2.89	3.05	3.04	3.16
310	training	209	1-bromo-2-nitrobenzene	577-19-5	O=N(=O)c(c(ccc1)Br)c1	2.420	2.59	2.69	2.47	2.68	2.70	2.71	2.52
311	test	98	1-bromo-3-nitrobenzene	585-79-5	O=N(=O)c(cccc1Br)c1	2.420	2.61	2.55	2.47	2.68	2.70	2.71	2.64
312	training	210	1-bromo-4-nitrobenzene	586-78-7	O=N(=O)c(ccc(c1)Br)c1	2.420	2.66	2.55	2.47	2.68	2.70	2.71	2.55
313	training	211	m-dibromobenzene	108-36-1	c(cccc1Br)(c1)Br	3.417	3.73	3.38	3.33	3.80	3.77	3.62	3.75
314	training	212	m-chloronitrobenzene	121-73-3	O=N(=O)c(cccc1Cl)c1	2.715	2.49	2.46	2.39	2.51	2.46	2.54	2.46
315	training	213	o-chloronitrobenzene	88-73-3	O=N(=O)c(c(ccc1)Cl)c1	2.596	2.48	2.46	2.39	2.51	2.46	2.54	2.24
316	test	99	p-chloronitrobenzene	100-00-5	N(=O)(=O)c(ccc(c1)Cl)c1	2.677	2.56	2.46	2.39	2.51	2.46	2.54	2.39
317	test	100	o-dichlorobenzene	95-50-1	c(c(ccc1)Cl)(c1)Cl	2.780	3.45	3.21	3.16	3.48	3.28	3.27	3.43
318	training	214	m-dichlorobenzene	541-73-1	c(cccc1Cl)(c1)Cl	2.780	3.45	3.21	3.16	3.48	3.28	3.27	3.53
319	training	215	p-dichlorobenzene	106-46-7	c(ccc(c1)Cl)(c1)Cl	2.780	3.46	3.21	3.16	3.48	3.28	3.27	3.44
320	test	101	2,3-dichlorophenol	576-24-9	Oc(c(c(cc1)Cl)Cl)c1	2.550	3.15	2.91	2.89	2.73	2.80	2.86	2.84
321	training	216	2,4-dichlorophenol	120-83-2	Oc(c(cc(c1)Cl)Cl)c1	2.550	3.14	2.91	2.89	2.73	2.80	2.86	3.06
322	training	217	2,5-dichlorophenol	583-78-8	Oc(c(ccc1Cl)Cl)c1	2.550	3.13	2.91	2.89	2.73	2.80	2.86	3.06
323	test	102	2,6-dichlorophenol	87-65-0	Oc(c(ccc1)Cl)c1Cl	2.550	3.15	2.91	2.89	2.73	2.80	2.86	2.75
324	training	218	3,4-dichlorophenol	95-77-2	Oc1ccc(Cl)c(Cl)c1	2.550	3.12	2.91	2.89	2.73	2.80	2.86	3.33
325	training	219	3,5-dichlorophenol	591-35-5	Oc1cc(Cl)cc(Cl)c1	2.550	3.09	2.91	2.89	2.73	2.80	2.86	3.62
326	test	103	m-difluorobenzene	372-18-9	Fc1cccc(F)c1	2.579	2.25	2.10	2.24	3.14	2.39	2.34	2.08
327	test	104	o-difluorobenzene	367-11-3	Fc1ccccc1F	2.666	2.24	2.10	2.24	3.14	2.39	2.34	2.37
328	training	220	p-difluorobenzene	540-36-3	Fc1ccc(F)cc1	2.536	2.26	2.10	2.24	3.14	2.39	2.34	1.93
329	test	105	m-dinitrobenzene	99-65-0	N(=O)(=O)c(cccc1N(=O)(=O))c1	2.188	1.70	1.72	1.62	1.89	1.63	1.81	1.49

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
330	training	221	o-dinitrobenzene	528-29-0	O=N(=O)c1ccccc1N(=O)=O	2.296	1.64	1.72	1.62	1.89	1.63	1.81	1.69
331	test	106	p-dinitrobenzene	100-25-4	N(=O)(=O)c(ccc(N(=O)(=O))c1)c1	2.171	1.70	1.72	1.62	1.89	1.63	1.81	1.46
332	training	222	bromobenzene	108-86-1	c(cccc1)(c1)Br	3.004	2.65	2.68	2.58	3.05	2.88	2.82	2.99
333	test	107	o-bromophenol	95-56-7	Oc(c(ccc1)Br)c1	2.410	2.52	2.38	2.31	2.30	2.40	2.41	2.35
334	training	223	m-bromophenol	591-20-8	Oc(cccc1Br)c1	2.410	2.46	2.38	2.31	2.30	2.40	2.41	2.63
335	training	224	p-bromophenol	106-41-2	Oc(ccc(c1)Br)c1	2.410	2.50	2.38	2.31	2.30	2.40	2.41	2.59
337	training	225	chlorobenzene	108-90-7	c(cccc1)(c1)Cl	2.220	2.78	2.59	2.49	2.88	2.64	2.64	2.89
338	training	226	m-chlorophenol	108-43-0	Oc(cccc1Cl)c1	1.820	2.35	2.30	2.23	2.13	2.16	2.24	2.50
339	training	227	o-chlorophenol	95-57-8	Oc(c(ccc1)Cl)c1	1.710	2.40	2.30	2.23	2.13	2.16	2.24	2.15
340	test	108	p-chlorophenol	106-48-9	Oc(ccc(c1)Cl)c1	1.850	2.37	2.30	2.23	2.13	2.16	2.24	2.39
341	training	228	3,4-dichloroaniline	95-76-1	Nc(ccc(c1Cl)Cl)c1	0.670	2.74	2.48	2.41	2.73	2.37	2.45	2.69
342	training	229	2,3-dichloroaniline	608-27-5	Nc(c(c(cc1)Cl)Cl)c1	0.670	2.73	2.48	2.41	2.73	2.37	2.45	2.86
343	test	109	2,4-dichloroaniline	554-00-7	Nc(c(cc(c1)Cl)Cl)c1	0.670	2.73	2.48	2.41	2.73	2.37	2.45	2.91
344	training	230	2,5-dichloroaniline	95-82-9	Nc(c(ccc1Cl)Cl)c1	0.670	2.72	2.48	2.41	2.73	2.37	2.45	2.92
345	training	231	2,6-dichloroaniline	608-31-1	Nc(c(ccc1)Cl)c1Cl	0.670	2.74	2.48	2.41	2.73	2.37	2.45	2.82
346	test	110	3,5-dichloroaniline	626-43-7	Nc(cc(cc1Cl)Cl)c1	0.670	2.71	2.48	2.41	2.73	2.37	2.45	2.90
347	training	232	fluorobenzene	462-06-6	Fc(cccc1)c1	2.612	2.18	2.04	2.04	2.70	2.19	2.18	2.27
348	training	233	iodobenzene	591-50-4	c(cccc1)(c1)I	3.161	3.00	2.91	2.41	3.21	3.16	3.09	3.25
349	training	234	nitrobenzene	98-95-3	N(=O)(=O)c(cccc1)c1	2.010	1.89	1.85	1.72	1.89	1.81	1.91	1.85
350	training	235	o-nitrophenol	88-75-5	N(=O)(=O)c(c(O)ccc1)c1	2.060	1.91	1.55	1.46	1.35	1.91	2.15	1.79
351	test	111	m-nitrophenol	554-84-7	N(=O)(=O)c(cccc1O)c1	1.720	1.92	1.55	1.46	1.35	1.91	1.51	2.00
352	training	236	p-nitrophenol	100-02-7	N(=O)(=O)c(ccc(O)c1)c1	2.720	1.93	1.55	1.46	1.35	1.91	1.51	1.91
353	test	112	benzene	71-43-2	c(cccc1)c1	1.870	2.03	1.98	1.83	2.26	1.99	2.02	2.13
354	training	237	o-bromoaniline	615-36-1	Nc(c(ccc1)Br)c1	1.960	2.14	1.96	1.83	2.30	1.97	2.00	2.11
355	training	238	m-bromoaniline	591-19-5	Nc(cccc1Br)c1	1.960	2.16	1.96	1.83	2.30	1.97	2.00	2.10
356	test	113	p-bromoaniline	106-40-1	Nc(ccc(c1)Br)c1	1.960	2.10	1.96	1.83	2.30	1.97	2.00	2.26
357	training	239	m-chloroaniline	108-42-9	Nc(cccc1Cl)c1	3.130	1.93	1.87	1.75	2.13	1.72	1.83	1.88
358	training	240	o-chloroaniline	95-51-2	Nc(c(ccc1)Cl)c1	3.130	1.93	1.87	1.75	2.13	1.72	1.83	1.90
359	test	114	p-chloroaniline	106-47-8	Nc(ccc(c1)Cl)c1	3.130	1.95	1.87	1.75	2.13	1.72	1.83	1.88
360	test	115	1α,2α,3ß,4α,5α,6ß-hexachlorocyclohexane	58-89-9	C(C(C(C(C1Cl)Cl)Cl)Cl)(C1Cl)Cl	3.410	3.94	3.90	4.16	4.09	4.26	3.83	3.80
361	test	116	1α,2ß,3α,4ß,5α,6ß-hexachlorocyclohexane	319-85-7	C(C(C(C(C1Cl)Cl)Cl)Cl)(C1Cl)Cl	3.120	3.94	3.90	4.16	4.09	4.26	3.83	3.80
362	training	241	1α,2α,3α,4ß,5α,6ß-hexachlorocyclohexane	319-86-8	C(C(C(C(C1Cl)Cl)Cl)Cl)(C1Cl)Cl	3.300	3.94	3.90	4.16	4.09	4.26	3.83	3.80
364	test	117	a-hexachlorocyclohexane	319-84-6	C(C(C(C(C1Cl)Cl)Cl)Cl)(C1Cl)Cl	3.300	3.94	3.90	4.16	4.09	4.26	3.83	3.80
366	test	118	m-nitroaniline	99-09-2	N(=O)(=O)c(cccc1N)c1	2.122	1.53	1.13	0.98	1.35	1.47	1.10	1.37
367	training	242	o-nitroaniline	88-74-4	N(=O)(=O)c(c(N)ccc1)c1	2.383	1.43	1.13	0.98	1.35	2.02	1.74	1.85
368	test	119	p-nitroaniline	100-01-6	N(=O)(=O)c(ccc(N)c1)c1	2.133	1.50	1.13	0.98	1.35	1.47	1.10	1.39
369	training	243	phenol	108-95-2	Oc(cccc1)c1	1.740	1.39	1.68	1.56	1.51	1.51	1.62	1.46
370	test	120	pyrocatechol	120-80-9	Oc(c(O)ccc1)c1	2.070	0.74	1.38	1.29	0.89	1.03	1.64	0.88
371	test	121	resorcinol	108-46-3	Oc(cccc1O)c1	1.020	0.70	1.38	1.29	0.89	1.03	1.21	0.80
372	test	122	p-hydroquinone	123-31-9	Oc(ccc(O)c1)c1	1.698	0.71	1.38	1.29	0.89	1.03	0.79	0.59
373	test	123	phenyl mercaptan	108-98-5	Sc(cccc1)c1	2.748	2.26	2.16	2.19	2.52	2.69	2.40	2.52
374	training	244	aniline	62-53-3	Nc(cccc1)c1	1.867	0.89	1.26	1.08	1.51	1.08	1.21	0.90

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
375	test	124	2-methylpyridine	109-06-8	n(c(ccc1)C)c1	1.981	1.25	1.32	0.96	0.85	1.35	1.09	1.11
376	training	245	3-methylpyridine	108-99-6	n(cccc1C)c1	2.030	1.11	1.22	1.17	0.85	1.35	1.00	1.20
377	test	125	4-methylpyridine	108-89-4	n(ccc(c1)C)c1	2.041	1.14	1.22	1.17	0.85	1.35	1.21	1.22
378	training	246	1,4-cyclohexadiene	628-41-1	C(=CCC=C1)C1	2.721	2.31	2.21	1.85	1.96	2.75	2.76	2.30
379	training	247	adiponitrile	111-69-3	C(#N)CCCCC(#N)	1.203	-0.14	2.33	1.05	0.45	0.35	0.23	-0.32
380	training	248	m-phenylenediamine	108-45-2	Nc(cccc1N)c1	1.197	0.01	0.54	0.34	0.89	-0.39	0.39	-0.33
381	training	249	o-phenylenediamine	95-54-5	Nc(c(N)ccc1)c1	1.459	-0.08	0.54	0.34	0.89	0.16	0.39	0.15
382	training	250	p-phenylenediamine	106-50-3	Nc(ccc(N)c1)c1	1.214	-0.01	0.54	0.34	0.89	-0.39	-0.03	-0.30
383	test	126	phenylhydrazine	100-63-0	N(N)c(cccc1)c1	2.057	0.95	1.50	0.94	1.44	0.79	0.78	1.25
384	test	127	2-ethylfuran	3208-16-0	C1=COC(CC)=C1	2.683	2.50	1.76	1.75	0.94	2.40	1.20	2.40
385	test	128	2-cyclohexen-1-one	930-68-7	O=C(C=CCC1)C1	1.709	0.97	1.64	1.15	0.94	1.20	0.77	0.61
386	training	251	5-hexyn-2-one	2550-28-9	C#CCCC(=O)C	1.693	1.13	1.13	1.52	1.34	0.47	0.44	0.58
387	test	129	ascorbic acid	50-81-7	O1C(=O)C(O)=C(O)C1C(O)CO	0.485	-1.58	-2.23	-1.76	-2.33	-1.88	-0.61	-1.64
388	training	252	citric acid	77-92-9	O=C(O)C(O)(CC(=O)O)CC(=O)O	0.441	-1.33	-2.15	-1.39	-1.17	-1.67	-2.25	-1.72
389	training	253	cyclohexene	110-83-8	C(=CCCC1)C1	2.933	2.77	2.31	2.29	2.07	2.96	2.90	2.86
390	test	130	1,5-hexadiene	592-42-7	C(=C)CCC=C	2.900	3.05	2.65	2.33	2.35	3.02	3.02	2.87
391	training	254	cis-2,trans-4-hexadiene	5194-50-3	CC=CC=CC	2.900	3.24	2.52	2.22	2.35	2.86	3.00	2.80
392	training	255	trans-2,trans-4-hexadiene	5194-51-4	CC=CC=CC	3.014	3.24	2.52	2.22	2.35	2.86	3.00	2.80
393	training	256	1-hexyne	693-02-7	C(#C)CCCC	2.862	2.63	2.23	3.29	2.46	2.52	2.62	2.42
394	training	257	cyclohexanone	108-94-1	O=C(CCCC1)C1	1.818	1.03	1.74	1.17	1.05	1.13	0.92	0.81
395	test	131	5-hexen-2-one	109-49-9	O=C(CCC=C)C	1.932	1.08	1.84	0.94	1.34	1.10	1.21	1.02
396	training	258	ethyl methacrylate	97-63-2	O=C(OCC)C(=C)C	2.432	1.69	1.56	1.51	1.27	1.77	1.10	1.94
397	training	259	ethylacetoacetate	141-97-9	O=C(OCC)CC(=O)C	1.513	0.19	0.74	0.14	0.43	-0.20	0.71	0.24
398	test	132	adipic acid	124-04-9	O=C(O)CCCCC(=O)O	1.421	0.13	0.53	0.50	0.43	0.23	0.01	0.08
399	test	133	diethyl oxalate	95-92-1	O=C(OCC)C(=O)OCC	1.682	1.15	0.45	0.72	0.43	0.40	0.80	0.56
400	training	260	bromocyclohexane	108-85-0	BrC(CCCC1)C1	3.118	3.63	2.71	2.85	2.73	3.45	2.98	3.20
401	training	261	hexanenitrile	628-73-9	N#CCCCCC	2.247	1.64	2.78	2.08	1.44	1.82	1.99	1.66
402	training	262	epsilon-caprolactam	105-60-2	O=C(NCCCC1)C1	1.274	-0.08	0.74	0.54	0.57	0.66	0.40	-0.10
403	training	263	cyclohexanone oxime	100-64-1	N(O)=C(CCCC1)C1	1.834	1.52	1.98	1.26	0.98	0.91	1.16	1.19
404	training	264	methylcyclopentane	96-37-7	C(CCC1)(C1)C	3.210	3.15	2.28	2.53	3.12	3.10	3.35	3.37
405	training	265	cyclohexane	110-82-7	C(CCCC1)C1	3.248	3.46	2.41	2.74	3.12	3.18	3.41	3.44
406	training	266	1-hexene	592-41-6	C(=C)CCCC	3.227	3.38	2.94	2.72	2.46	3.15	3.39	3.39
407	test	134	4-methyl-1-pentene	691-37-2	C(=C)CC(C)C	2.737	3.08	2.82	2.51	2.46	3.08	3.33	2.66
408	training	267	thiram	137-26-8	N(C(=S)SSC(N(C)C)=S)(C)C	3.010	2.18	-0.10	3.16	0.45	1.70	0.18	1.69
409	training	268	cyclohexanol	108-93-0	OC(CCCC1)C1	2.046	1.35	1.58	1.50	1.20	1.64	1.33	1.23
410	training	269	hexanal	66-25-1	O=CCCCCC	2.345	2.37	2.29	1.85	1.44	1.80	2.29	1.78
411	test	135	2-hexanone	591-78-6	O=C(CCCC)C	2.128	1.45	2.14	1.34	1.44	1.24	1.58	1.38
412	training	270	3-methyl-2-pentanone	565-61-7	CC(=O)C(C)CC	0.630	1.48	2.01	1.34	1.44	1.16	1.30	1.35
413	test	136	4-methyl-2-pentanone	108-10-1	O=C(CC(C)C)C	0.630	1.31	2.01	1.13	1.44	1.16	1.51	1.31
414	training	271	hexanoic acid	142-62-1	O=C(O)CCCCC	2.421	1.88	1.88	1.81	1.37	2.05	1.88	1.92
415	test	137	2-ethyl butyric acid	88-09-5	O=C(O)C(CC)CC	2.291	1.74	1.76	1.81	1.37	1.98	1.60	1.68
416	training	272	butyl acetate	123-86-4	O=C(OCCCC)C	2.367	1.84	1.84	1.35	1.37	1.85	1.59	1.78

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
417	training	273	isobutyl acetate	110-19-0	O=C(OCC(C)C)C	2.345	1.74	1.72	1.21	1.37	1.77	1.32	1.78
418	training	274	sec-butyl acetate	105-46-4	O=C(OC(CC)C)C	2.313	1.97	1.78	1.27	1.37	1.77	1.48	1.72
419	training	275	tert-butyl acetate	540-88-5	O=C(OC(C)(C)C)C	2.334	1.88	1.47	0.95	1.37	1.74	1.54	1.76
420	test	138	hydroxycaproic acid	1191-25-9	O=C(O)CCCCCO	1.818	0.29	0.93	0.58	0.54	0.59	0.12	-0.41
421	training	276	paraldehyde	123-63-7	O(C(OC(O1)C)C)C1C	1.741	0.33	1.08	0.09	0.70	0.70	1.22	0.67
422	training	277	glucose	50-99-7	OCC1C(O)C(O)C(O)C(O)O1	-0.386	-2.57	-2.32	-2.51	-2.48	-2.89	-2.34	-2.59
423	training	278	1-bromohexane	111-25-1	BrCCCCCC	3.444	3.88	3.46	3.23	3.13	3.63	3.80	3.80
424	training	279	cyclohexylamine	108-91-8	NC(CCCC1)C1	2.188	1.30	1.04	1.21	1.20	1.63	1.26	1.49
425	test	139	hexane	110-54-3	C(CCCC)C	3.553	4.02	3.23	3.11	3.52	3.29	3.75	3.90
426	test	140	2,2-dimethylbutane	75-83-2	C(CC)(C)(C)C	3.455	3.74	3.16	2.65	3.52	3.18	3.73	3.03
427	test	141	2,3-dimethylbutane	79-29-8	C(C(C)C)(C)C	3.471	2.84	2.99	2.70	3.52	3.14	3.63	3.42
428	training	280	3-methylpentane	96-14-0	C(CC)(CC)C	3.335	3.98	3.11	2.90	3.52	3.21	3.69	3.18
429	training	281	lysine	56-87-1	O=C(O)C(N)CCCCN	-0.282	-3.76	-1.16	-0.68	-2.48	-2.99	-2.95	-3.05
430	training	282	1-hexanol	111-27-3	OCCCCCC	2.481	2.03	2.28	1.88	1.59	1.82	1.99	2.03
431	test	142	2-hexanol	626-93-7	OC(CCCC)C	2.334	1.75	2.22	1.80	1.59	1.75	1.88	1.76
432	training	283	3-hexanol	623-37-0	OC(CCC)CC	2.275	1.76	2.22	1.87	1.59	1.75	1.67	1.65
433	test	143	3,3-dimethyl-2-butanol	464-07-3	OC(C(C)(C)C)C	2.182	1.75	2.15	1.49	1.59	1.64	1.43	1.48
434	test	144	dipropyl ether	111-43-3	O(CCC)CCC	2.481	2.04	2.24	1.80	1.59	2.03	1.58	2.03
435	test	145	diisopropyl ether	108-20-3	O(C(C)C)C(C)C	2.204	1.69	2.12	1.50	1.59	1.88	1.79	1.52
436	training	284	ethyl butyl ether	628-81-9	O(CCCC)CC	2.481	2.10	2.24	1.73	1.59	2.03	1.79	2.03
437	test	146	acetal	105-57-7	O(C(OCC)C)CC	1.834	1.19	1.23	0.78	1.11	1.20	1.27	0.84
438	training	285	2-butoxyethanol	111-76-2	O(CCCC)CCO	1.829	0.78	1.29	0.84	0.71	0.57	0.66	0.83
439	training	286	dipropyl sulfone	598-03-8	O=S(=O)(CCC)CCC	1.589	0.36	1.34	1.52	1.27	0.86	1.52	1.39
440	training	287	diethylene glycol dimethyl ether	111-96-6	O(CCOC)CCOC	1.181	0.12	0.31	-0.21	-0.13	-0.48	-0.36	-0.36
441	training	288	2-(2-ethoxyethoxy)ethanol	111-90-0	O(CCOCC)CCO	1.083	-0.16	0.30	-0.27	-0.13	-0.69	-0.46	-0.54
442	training	289	trimethylolpropane	77-99-6	OCC(CC)(CO)CO	0.572	-0.76	0.30	-0.62	-0.13	0.19	-0.72	-0.82
443	test	147	sorbitol	50-70-4	OCC(O)C(O)C(O)C(O)CO	0.180	-2.68	-2.73	-2.94	-2.50	-3.01	-3.90	-3.10
444	test	148	hexylamine	111-26-2	NCCCCCC	2.498	1.98	1.74	1.59	1.59	1.82	1.92	2.06
445	training	290	di-propylamine	142-84-7	N(CCC)CCC	2.285	1.74	1.77	1.53	1.59	1.79	1.39	1.67
446	training	291	diisopropylamine	108-18-9	N(C(C)C)C(C)C	2.139	1.12	1.65	1.24	1.59	1.64	1.59	1.40
447	test	149	triethylamine	121-44-8	N(CC)(CC)CC	2.166	1.57	1.50	1.37	1.59	1.51	1.33	1.45
448	test	150	diisopropanolamine	110-97-4	OC(C)CNCC(O)C	0.931	-0.40	-0.25	-0.54	-0.13	-0.88	-0.67	-0.82
449	training	292	triethanolamine	102-71-6	OCCN(CCO)CCO	0.833	-1.38	-1.36	-1.30	-0.94	-2.48	-2.06	-1.00
450	training	293	triethyl phosphate	78-40-0	O=P(OCC)(OCC)OCC	1.812	0.71	0.80	0.69	1.35	0.87	0.58	0.80
451	training	294	hexamethyl phosphoramide	680-31-9	O=P(N(C)C)(N(C)C)N(C)C	1.529	0.03	-0.05	0.46	1.35	-0.22	-0.56	0.28
452	test	151	hexamethyldisiloxane	107-46-0	C[Si](C)(C)O[Si](C)(C)C	3.662	2.89	3.49	3.50	1.59	4.76	2.92	2.92
453	test	152	3-nitrobenzotrifluoride	98-46-4	O=N(=O)c(cccc1C(F)(F)F)c1	2.802	2.55	2.75	2.67	3.00	2.77	2.84	2.62
454	training	295	2-bromobenzoic acid	88-65-3	O=C(O)c(c(ccc1)Br)c1	2.574	2.54	2.20	2.18	2.46	2.42	2.43	2.20
455	training	296	3-bromobenzoic acid	585-76-2	O=C(O)c(cccc1Br)c1	2.938	2.42	2.20	2.18	2.46	2.76	2.44	2.87
456	training	297	4-bromobenzoic acid	586-76-5	O=C(O)c(ccc(c1)Br)c1	2.933	2.43	2.20	2.18	2.46	2.76	2.44	2.86
457	training	298	o-chlorobenzoic acid	118-91-2	O=C(O)c(c(ccc1)Cl)c1	2.492	2.39	2.12	2.10	2.30	2.18	2.26	2.05
458	training	299	chloramben	133-90-4	O=C(O)c(c(c(N)cc1Cl)Cl)c1	1.250	2.05	2.01	2.02	2.32	1.90	2.06	2.02

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
459	test	153	4,5,6-trichloroguaiacol	2668-24-8	COc1cc(Cl)c(Cl)c(Cl)c1O	2.990	3.81	3.42	3.54	2.74	3.27	3.40	3.79
460	test	154	benzotrifluoride	98-08-8	FC(F)(F)c(cccc1)c1	3.036	2.91	2.74	2.77	3.37	2.96	2.95	3.01
461	test	155	benzonitrile	100-47-0	C(#N)c(cccc1)c1	2.226	1.55	1.79	1.71	1.77	1.54	1.75	1.56
462	test	156	benzothiazole	95-16-9	c1ccc2ncsc2c1	2.470	2.13	2.00	1.91	1.50	2.17	1.94	2.01
463	test	157	2,4,6-trinitrotoluene	118-96-7	N(=O)(=O)c(cc(N(=O)(=O))c(c1N(=O)(=O))C)c1	2.247	1.50	2.31	2.00	2.36	1.99	1.92	1.60
464	test	158	2,4-dichlorotoluene	95-73-8	c(ccc(c1Cl)C)(c1)Cl	3.684	3.95	3.52	3.64	3.80	3.83	3.49	4.24
465	training	300	3,4-dichlorophenyl urea	2327-02-8	NC(=O)Nc1ccc(Cl)c(Cl)c1	2.490	2.35	2.08	2.07	2.05	2.00	1.64	2.64
466	training	301	2-(trifluoromethyl)aniline	88-17-5	FC(F)(F)c(c(N)ccc1)c1	2.360	2.24	2.02	2.03	2.62	2.04	2.13	2.41
467	test	159	3-(trifluoromethyl)aniline	98-16-8	FC(F)(F)c(cccc1N)c1	2.360	2.23	2.02	2.03	2.62	2.04	2.13	2.29
468	training	302	4-(trifluoromethyl)aniline	455-14-1	FC(F)(F)c(ccc(N)c1)c1	2.360	2.30	2.02	2.03	2.62	2.04	2.13	2.39
469	training	303	3-(trifluoromethoxy)aniline	1535-73-5	Nc1cc(OC(F)(F)F)ccc1	2.360	2.37	2.09	3.20	1.74	2.12	2.37	2.14
470	test	160	2-(trifluoromethoxy)aniline	1535-75-7	Nc1c(OC(F)(F)F)cccc1	2.360	2.45	2.09	3.20	1.74	2.12	2.37	2.39
471	training	304	4-(trifluoromethoxy)aniline	461-82-5	NC1=CC=C(OC(F)(F)F)C=C1	2.360	2.34	2.09	3.20	1.74	2.12	2.37	2.16
472	training	305	3-(trifluoromethylthio)aniline	369-68-6	Nc1cccc(c1)SC(F)(F)F	2.360	2.88	3.78	3.76	2.62	2.64	2.95	2.66
473	test	161	4-(trifluoromethylthio)aniline	372-16-7	NC1=CC=C(SC(F)(F)F)C=C1	2.360	2.89	3.78	3.76	2.62	2.64	2.95	2.96
474	training	306	1H-benzimidazole	51-17-2	n(c(c(n1)ccc2)c2)c1	2.106	1.67	1.62	1.35	1.43	1.23	1.32	1.32
475	training	307	2-hydroxybenzimidazole	615-16-7	O=C(Nc(c1ccc2)c2)N1	1.986	0.74	1.44	0.69	0.86	1.20	0.46	1.12
476	training	308	2,4-dinitrotoluene	121-14-2	N(=O)(=O)c(ccc(c1N(=O)(=O))C)c1	2.454	1.90	2.04	2.11	2.24	2.18	2.03	1.98
477	training	309	2,6-dinitrotoluene	606-20-2	N(=O)(=O)c(c(c(N(=O)(=O))cc1)C)c1	2.519	1.81	2.31	2.11	2.24	2.18	2.03	2.10
478	test	162	3,4-dinitrotoluene	610-39-9	O=N(=O)c(c(N(=O)=O)ccc1C)c1	2.509	1.89	2.31	2.11	2.24	2.18	2.24	2.08
479	training	310	benzaldehyde	100-52-7	O=Cc(cccc1)c1	2.182	1.60	1.67	1.59	1.77	1.71	1.72	1.48
480	training	311	benzoic acid	65-85-0	O=C(O)c(cccc1)c1	1.950	1.72	1.50	1.43	1.70	1.87	1.64	1.87
481	training	312	p-hydroxybenzaldehyde	123-08-0	O=Cc(ccc(O)c1)c1	2.111	1.27	1.37	1.32	1.16	1.23	1.31	1.35
482	training	313	salicylaldehyde	90-02-8	O=Cc(c(O)ccc1)c1	2.362	1.22	1.37	1.32	1.67	2.01	1.74	1.81
483	training	314	1,3-benzodioxole	274-09-9	O(c(c(O1)ccc2)c2)C1	2.509	1.71	2.08	1.60	1.26	2.05	1.78	2.08
484	training	315	phenyl formate	1864-94-4	O=COc1ccccc1	2.062	1.31	1.64	1.55	1.70	1.04	1.57	1.26
485	test	163	salicylic acid	69-72-7	O=C(O)c(c(O)ccc1)c1	2.574	1.96	1.20	1.17	1.64	2.24	2.43	2.26
486	test	164	p-bromotoluene	106-38-7	c(ccc(c1)Br)(c1)C	3.237	3.35	2.99	3.07	3.37	3.43	3.26	3.39
487	test	165	(bromomethyl)benzene	100-39-0	BrCc(cccc1)c1	2.965	2.76	2.60	2.58	3.10	2.88	2.82	2.92
488	training	316	(4-bromophenyl)urea	1967-25-5	NC(=O)Nc1ccc(Br)cc1	2.120	2.10	1.55	1.49	1.63	1.60	1.20	1.98
489	training	317	benzyl chloride	100-44-7	c(cccc1)(c1)CCl	2.628	2.51	2.55	2.43	2.94	2.79	2.49	2.30
490	training	318	o-chlorotoluene	95-49-8	c(c(ccc1)Cl)(c1)C	3.237	3.27	2.91	2.98	3.21	3.18	2.87	3.42
491	test	166	p-chlorotoluene	106-43-4	c(ccc(c1)Cl)(c1)C	3.189	3.30	2.91	2.98	3.21	3.18	3.08	3.33
492	training	319	2-chlorophenyl urea	114-38-5	NC(=O)Nc1ccccc1Cl	1.610	1.74	1.47	1.40	1.47	1.35	1.02	1.27
493	test	167	3-chlorophenyl urea	1967-27-7	NC(=O)Nc1cccc(Cl)c1	2.010	1.58	1.47	1.40	1.47	1.35	1.02	1.82
494	training	320	p-fluorotoluene	352-32-9	Fc(ccc(c1)C)c1	2.781	2.65	2.36	2.52	3.05	2.74	2.62	2.80
495	training	321	2-fluorophenyl urea	656-31-5	NC(=O)Nc1ccccc1F	1.310	1.20	0.92	0.95	1.30	0.91	0.56	0.88
496	training	322	3-fluorophenyl urea	770-19-4	NC(=O)Nc1cccc(F)c1	1.770	1.07	0.92	0.95	1.30	0.91	0.56	1.29
497	training	323	4-fluorophenyl urea	659-30-3	NC(=O)Nc1ccc(F)cc1	1.520	1.13	0.92	0.95	1.30	0.91	0.56	1.04
498	test	168	formanilide	103-70-8	O=CNc(cccc1)c1	2.003	1.20	1.32	0.90	1.29	1.13	1.04	1.15
499	test	169	m-nitrotoluene	99-08-1	N(=O)(=O)c(cccc1C)c1	2.710	2.32	2.17	2.21	2.24	2.36	2.35	2.42
500	test	170	o-nitrotoluene	88-72-2	N(=O)(=O)c(c(ccc1)C)c1	2.628	2.32	2.17	2.21	2.24	2.36	2.14	2.30

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
501	training	324	p-nitrotoluene	99-99-0	N(=O)(=O)c(ccc(c1)C)c1	2.693	2.34	2.17	2.21	2.24	2.36	2.35	2.30
502	test	171	o-nitroanisole	91-23-6	O=N(=O)c(c(OC)ccc1)c1	2.318	2.02	1.75	1.71	1.70	1.89	1.83	1.73
503	training	325	4-methyl-3-nitrophenol	2042-14-0	c1c(O)ccc(C)c1N(=O)=O	2.610	2.33	2.00	1.94	1.70	2.46	1.73	2.07
504	test	172	3-methyl-4-nitrophenol	2581-34-2	O=N(=O)c(c(cc(O)c1)C)c1	2.610	2.27	2.00	1.94	1.70	2.46	1.73	2.48
505	training	326	3-methyl-2-nitrophenol	4920-77-8	c1c(C)c(N(=O)=O)c(O)cc1	2.610	2.28	2.00	1.94	1.70	2.46	2.38	2.29
506	training	327	2-methyl-3-nitrophenol	5460-31-1	N(=O)(=O)c1c(C)c(O)ccc1	2.610	2.28	2.00	1.94	1.70	2.46	1.73	2.11
507	test	173	5-methyl-2-nitrophenol	700-38-9	Cc1ccc(N(=O)=O)c(O)c1	2.610	2.29	2.00	1.94	1.70	2.46	2.59	2.31
508	training	328	toluene	108-88-3	c(cccc1)(c1)C	1.970	2.56	2.30	2.32	2.61	2.54	2.46	2.73
509	training	329	2-bromo-4-methylaniline	583-68-6	Nc(c(cc(c1)C)Br)c1	1.960	2.52	2.27	2.32	2.62	2.51	2.44	2.00
510	training	330	2-bromo-5-methylaniline	53078-85-6	CC1=CC(=C(C=C1)Br)N	1.960	2.51	2.27	2.32	2.62	2.51	2.44	2.00
511	test	174	3-bromo-4-methylaniline	7745-91-7	Nc1cc(Br)c(C)cc1	1.960	2.54	2.27	2.32	2.62	2.51	2.23	1.51
512	training	331	4-bromo-2-methylaniline	583-75-5	Nc(c(cc(c1)Br)C)c1	1.960	2.51	2.27	2.32	2.62	2.51	2.23	1.95
513	training	332	4-bromo-3-methylaniline	6933-10-4	Cc1cc(N)ccc1Br	1.960	2.55	2.27	2.32	2.62	2.51	2.23	2.94
514	test	175	5-bromo-2-methylaniline	39478-78-9	Nc1cc(Br)ccc1C	1.960	2.50	2.27	2.32	2.62	2.51	2.23	1.51
515	training	333	3-bromo-2-methylaniline	55289-36-6	Brc1cccc(N)c1C	1.960	2.50	2.27	2.32	2.62	2.51	2.23	2.30
516	training	334	3-chloroanisidine	5345-54-0	COc1ccc(N)cc1Cl	1.930	1.91	1.77	1.73	1.85	1.80	1.74	2.04
517	training	335	phenylurea	64-10-8	O=C(Nc(cccc1)c1)N	1.350	0.85	0.86	0.74	0.86	0.71	0.40	0.83
518	training	336	anisole	100-66-3	O(c(cccc1)c1)C	2.525	2.10	1.88	1.81	1.86	2.07	1.94	2.11
519	test	176	benzyl alcohol	100-51-6	OCc(cccc1)c1	1.948	1.07	1.36	1.23	1.59	1.08	1.21	1.10
520	test	177	m-cresol	108-39-4	Oc(cccc1C)c1	1.540	1.93	2.00	2.05	1.86	2.06	2.05	1.96
521	test	178	o-cresol	95-48-7	Oc(c(ccc1)C)c1	1.340	1.89	2.00	2.05	1.86	2.06	1.84	1.95
522	training	337	p-cresol	106-44-5	Oc(ccc(c1)C)c1	1.690	1.95	2.00	2.05	1.86	2.06	2.05	1.94
524	test	179	guaiacol	90-05-1	O(c(c(O)ccc1)c1)C	1.600	1.32	1.58	1.55	1.25	1.34	1.53	1.32
525	test	180	p-methoxyphenol	150-76-5	O(c(ccc(O)c1)c1)C	1.750	1.31	1.58	1.55	1.25	1.59	1.53	1.34
526	training	338	3-methoxyphenol	150-19-6	O(c(cccc1O)c1)C	1.550	1.32	1.58	1.55	1.25	1.59	1.53	1.58
527	test	181	benzylamine	100-46-9	NCc(cccc1)c1	1.970	0.90	0.82	0.94	1.59	1.07	1.14	1.09
528	test	182	N-methylaniline	100-61-8	N(c(cccc1)c1)C	2.280	1.68	1.55	1.64	1.86	1.62	1.67	1.66
529	training	339	m-toluidine	108-44-1	Nc(cccc1C)c1	1.740	1.32	1.57	1.57	1.86	1.62	1.64	1.40
530	test	183	o-toluidine	95-53-4	Nc(c(ccc1)C)c1	1.740	1.32	1.57	1.57	1.86	1.62	1.43	1.32
531	training	340	p-toluidine	106-49-0	Nc(ccc(c1)C)c1	1.900	1.34	1.57	1.57	1.86	1.62	1.64	1.39
532	training	341	2,6-dimethylpyridine	108-48-5	n(c(ccc1)C)c1C	2.291	1.60	1.74	1.25	1.20	1.90	1.40	1.68
533	test	184	m-toluenediamine	95-80-7	Nc(c(ccc1N)C)c1	1.453	0.37	0.85	0.82	1.25	0.16	0.61	0.14
534	test	185	simazine	122-34-9	n(c(nc(n1)NCC)NCC)c1Cl	2.080	2.48	2.07	2.16	2.27	2.40	1.20	2.18
535	training	342	butyl acrylate	141-32-2	O=C(OCCCC)C=C	2.661	2.20	2.01	2.05	1.62	2.20	1.88	2.36
536	training	343	isobutyl acrylate	106-63-8	O=C(OCC(C)C)C=C	2.585	2.14	1.89	1.91	1.62	2.13	1.61	2.22
537	training	344	diethyl malonate	105-53-3	O=C(OCC)CC(=O)OCC	1.899	0.93	0.91	0.75	0.79	0.90	1.19	0.96
538	test	186	oxamyl	23135-22-0	CNC(=O)ON=C(SC)C(=O)N(C)C	0.900	-0.16	-0.22	0.21	-0.46	-1.20	0.33	-0.47
539	training	345	mevinphos	7786-34-7	COC(=O)C=C(C)OP(=O)(OC)OC	1.640	0.71	0.67	0.02	0.75	-0.24	0.04	1.20
540	training	346	methylcyclohexane	108-87-2	C(CCCC1)(C1)C	3.488	3.90	2.60	2.99	3.48	3.59	3.92	3.61
541	training	347	cycloheptane	291-64-5	C1CCCCCC1	3.553	4.01	2.73	3.19	3.48	3.67	3.98	4.00
542	training	348	1-heptene	592-76-7	C(=C)CCCCC	3.548	4.00	3.40	3.17	2.81	3.64	3.96	3.99
543	training	349	aldicarb	116-06-3	O=C(ON=CC(SC)(C)C)NC	1.300	1.58	2.44	1.17	0.79	1.36	1.68	1.13

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
544	test	187	2-heptanone	110-43-0	O=C(CCCCC)C	2.454	1.92	2.60	1.79	1.79	1.73	2.14	1.98
545	training	350	5-methyl-2-hexanone	110-12-3	O=C(CCC(C)C)C	2.400	1.88	2.48	1.59	1.79	1.66	2.08	1.88
546	training	351	2,4-dimethyl-3-pentanone	565-80-0	O=C(C(C)C)C(C)C	2.389	1.91	2.35	2.02	1.79	1.58	1.28	1.86
547	training	352	cis-2-methylcyclohexanol	7443-70-1	C1C(C)C(O)CCC1	2.378	1.80	1.77	1.82	1.55	2.05	1.63	1.84
548	training	353	trans-2-methylcyclohexanol	7443-52-9	C1C(C)C(O)CCC1	2.367	1.80	1.77	1.82	1.55	2.05	1.63	1.84
549	test	188	heptanoic acid	111-14-8	O=C(O)CCCCCC	2.693	2.41	2.34	2.26	1.73	2.54	2.45	2.51
550	training	354	1-bromoheptane	629-04-9	BrCCCCCCC	3.749	4.40	3.93	3.69	3.45	4.12	4.37	4.36
551	training	355	1-chloroheptane	629-06-1	ClCCCCCCC	3.635	4.30	3.88	3.54	3.29	4.03	4.03	4.15
552	training	356	heptane	142-82-5	C(CCCCC)C	3.825	4.33	3.70	3.57	3.87	3.78	4.32	4.44
553	training	357	1-heptanol	111-70-6	OCCCCCCC	2.802	2.53	2.74	2.34	1.94	2.31	2.56	2.72
554	training	358	2-heptanol	543-49-7	OC(CCCCC)C	2.634	2.34	2.68	2.26	1.94	2.24	2.45	2.31
555	training	359	3-heptanol	589-82-2	OC(CCCC)CC	2.596	2.29	2.68	2.33	1.94	2.24	2.24	2.24
556	test	189	4-heptanol	589-55-9	OC(CCC)CCC	2.585	2.26	2.68	2.33	1.94	2.24	2.24	2.22
557	training	360	heptylamine	111-68-2	NCCCCCCC	2.775	2.57	2.21	2.05	1.94	2.31	2.49	2.57
558	test	190	phorate	298-02-2	CCOP(=S)(OCC)SCSCC	2.820	3.71	3.80	2.93	1.32	3.37	3.81	3.56
559	training	361	chlorothalonil	1897-45-6	N#Cc(c(c(c(c1C#N)Cl)Cl)Cl)c1Cl	2.980	3.98	4.06	4.24	3.13	3.66	3.96	2.90
560	training	362	phthalic anhydride	85-44-9	O=C(OC(=O)c1cccc2)c12	2.247	0.89	1.17	1.30	2.19	2.07	1.52	1.29
561	training	363	ethynylbenzene	536-74-3	C(c(cccc1)c1)#C	2.683	2.50	2.29	2.96	2.85	2.26	2.52	2.53
562	training	364	3,6-dichloro-2-methoxybenzoic acid	1918-00-9	COc1c(Cl)ccc(Cl)c1C(O)=O	0.990	2.65	2.62	2.75	2.62	2.14	2.80	2.21
563	training	365	(2,4-dichlorophenoxy)acetic acid	94-75-7	O=C(O)COc(c(cc(c1)Cl)Cl)c1	2.110	2.82	2.19	2.81	2.35	2.62	2.56	2.81
564	training	366	quinoxaline	91-19-0	n(c(c(nc1)ccc2)c2)c1	1.965	1.12	1.42	1.29	0.87	1.12	1.45	1.32
565	test	191	benzofuran	271-89-6	c1ccc2ccoc2c1	2.829	2.75	2.37	2.13	1.83	2.54	2.07	2.67
566	training	367	isophthalic acid	121-91-5	O=C(O)c(cccc1C(=O)O)c1	2.280	1.04	1.02	1.04	1.38	1.76	1.25	1.66
567	training	368	phthalic acid	88-99-3	O=C(O)c(c(ccc1)C(=O)O)c1	1.774	1.22	1.02	1.04	1.38	1.07	1.25	0.73
568	training	369	terephthalic acid	100-21-0	O=C(O)c(ccc(c1)C(=O)O)c1	2.465	1.01	1.02	1.04	1.38	1.76	1.25	2.00
569	training	370	benzothiophene	95-15-8	s(c(c(c1)ccc2)c2)c1	3.074	3.24	2.86	2.69	2.85	2.99	2.80	3.12
570	training	371	3-(trifluoromethylphenyl) urea	13114-87-9	O=C(Nc(cccc1C(F)(F)F)c1)N	1.600	1.73	1.62	1.68	1.94	1.67	1.32	2.31
571	training	372	indole	120-72-9	c1ccc2ccnc2c1	2.541	2.29	2.05	2.12	1.83	2.05	2.17	2.14
572	training	373	benzeneacetonitrile	140-29-4	C(#N)Cc(cccc1)c1	2.226	1.42	2.20	1.74	1.83	1.56	1.56	1.56
573	training	374	styrene	100-42-5	c(cccc1)(c1)C=C	3.036	2.92	2.58	2.38	2.85	2.89	2.96	2.95
574	training	375	1,3,5,7-cyclooctatetraene	629-20-9	C1=CC=CC=CC=C1	3.053	3.10	2.78	1.87	2.46	3.30	2.78	3.08
575	training	376	acetophenone	98-86-2	O=C(c(cccc1)c1)C	2.264	1.65	1.90	1.57	2.10	1.67	1.86	1.58
576	test	192	benzeneacetaldehyde	122-78-1	O=CCc(cccc1)c1	2.345	1.75	1.71	1.52	1.83	1.54	1.86	1.78
577	training	377	2-methylbenzaldehyde	529-20-4	Cc1ccccc1C=O	2.606	1.91	1.98	2.08	2.10	2.26	2.15	2.09
578	training	378	2,3-dihydrobenzofuran	496-16-2	O(c(c(ccc1)C2)c1)C2	2.541	2.16	2.27	1.89	1.80	2.51	1.78	2.14
579	training	379	phenyloxirane	96-09-3	O(C1c(cccc2)c2)C1	2.253	1.72	1.56	1.48	1.53	1.59	1.62	1.61
580	training	380	methyl benzoate	93-58-3	O=C(OC)c(cccc1)c1	2.574	1.98	1.95	1.68	2.03	1.83	1.96	2.12
581	training	381	o-toluic acid	118-90-1	O=C(O)c(c(ccc1)C)c1	2.639	2.03	1.82	1.92	2.03	2.08	2.07	2.46
582	training	382	p-toluic acid	99-94-5	O=C(O)c(ccc(c1)C)c1	2.650	2.12	1.82	1.92	2.03	2.42	2.07	2.27
583	training	383	benzeneacetic acid	103-82-2	O=C(O)Cc(cccc1)c1	2.144	1.72	1.30	1.47	1.77	1.43	1.45	1.41
584	training	384	phenyl acetate	122-79-2	O=C(Oc(cccc1)c1)C	2.188	1.59	1.91	1.60	2.03	1.59	1.81	1.49
585	training	385	m-toluic acid	99-04-7	O=C(O)c(cccc1C)c1	2.666	2.08	1.82	1.92	2.03	2.42	2.07	2.37

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
586	test	193	methyl salicylate	119-36-8	O=C(OC)c(c(O)ccc1)c1	2.764	2.07	1.66	1.42	1.98	2.60	2.54	2.34
587	training	386	vanillin	121-33-5	O=Cc(ccc(O)c1OC)c1	2.035	1.31	1.27	1.31	0.92	1.05	1.23	1.21
588	test	194	acetanilide	103-84-4	O=C(Nc(cccc1)c1)C	2.008	1.05	1.59	0.95	1.63	1.10	1.28	1.16
589	test	195	ethylbenzene	100-41-4	c(cccc1)(c1)CC	2.730	3.27	2.65	2.77	2.94	3.03	2.92	3.15
590	training	387	o-xylene	95-47-6	c(c(ccc1)C)(c1)C	2.700	3.16	2.61	2.80	2.94	3.09	2.90	3.12
591	test	196	m-xylene	108-38-3	c(cccc1C)(c1)C	2.460	3.15	2.61	2.80	2.94	3.09	2.90	3.20
592	training	388	p-xylene	106-42-3	c(ccc(c1)C)(c1)C	2.770	3.15	2.61	2.80	2.94	3.09	2.90	3.15
594	training	389	methyl parathion	298-00-0	COP(=S)(OC)Oc1ccc(cc1)N(=O)=O	2.640	2.97	2.78	2.20	1.60	2.75	3.07	2.86
596	training	390	phenetole	103-73-1	O(c(cccc1)c1)CC	2.742	2.56	2.31	2.16	2.19	2.57	2.36	2.51
597	training	391	2-phenylethanol	60-12-8	OCCc(cccc1)c1	2.117	1.51	1.70	1.55	1.92	1.57	1.37	1.36
598	training	392	3-methylbenzenemethanol	587-03-1	Cc1cccc(CO)c1	2.247	1.53	1.67	1.71	1.92	1.62	1.65	1.60
599	training	393	4-methylbenzenemethanol	589-18-4	OCc(ccc(c1)C)c1	2.237	1.54	1.67	1.71	1.92	1.62	1.65	1.58
600	training	394	1-phenylethanol	98-85-1	OC(c(cccc1)c1)C	1.570	1.58	1.82	1.60	1.92	1.49	1.74	1.42
601	test	197	o-ethylphenol	90-00-6	Oc(c(ccc1)CC)c1	2.721	2.45	2.35	2.51	2.19	2.55	2.31	2.47
602	training	395	m-ethylphenol	620-17-7	Oc1cc(CC)ccc1	2.737	2.53	2.36	2.51	2.19	2.55	2.52	2.40
603	training	396	p-ethylphenol	123-07-9	Oc(ccc(c1)CC)c1	2.737	2.54	2.36	2.51	2.19	2.55	2.52	2.58
604	test	198	2,3-xylenol	526-75-0	Oc(c(c(cc1)C)C)c1	2.660	2.34	2.31	2.54	2.19	2.61	2.28	2.65
605	training	397	2,4-xylenol	105-67-9	Oc(c(cc(c1)C)C)c1	2.660	2.37	2.31	2.54	2.19	2.61	2.28	2.30
606	training	398	2,5-xylenol	95-87-4	Oc(c(ccc1C)C)c1	2.660	2.35	2.31	2.54	2.19	2.61	2.28	2.33
607	test	199	2,6-xylenol	576-26-1	Oc(c(ccc1)C)c1C	2.660	2.32	2.31	2.54	2.19	2.61	2.07	2.36
608	training	399	3,4-xylenol	95-65-8	Oc(ccc(c1C)C)c1	2.660	2.41	2.31	2.54	2.19	2.61	2.49	2.23
609	training	400	3,5-xylenol	108-68-9	Oc(cc(cc1C)C)c1	2.660	2.38	2.31	2.54	2.19	2.61	2.49	2.35
610	training	401	benzyl methyl ether	538-86-3	O(Cc(cccc1)c1)C	2.111	1.69	1.81	1.63	1.92	1.77	1.73	1.92
611	training	402	2-methylanisole	578-58-5	O(c(c(ccc1)C)c1)C	2.868	2.60	2.19	2.30	2.19	2.62	2.16	2.74
612	training	403	3-methylanisole	100-84-5	O(c(cccc1C)c1)C	2.824	2.63	2.19	2.30	2.19	2.62	2.37	2.66
613	training	404	4-methylanisole	104-93-8	O(c(ccc(c1)C)c1)C	2.906	2.63	2.19	2.30	2.19	2.62	2.37	2.66
614	training	405	1,2-dimethoxybenzene	91-16-7	O(c(c(OC)ccc1)c1)C	2.509	2.10	1.77	1.80	1.58	1.64	1.58	1.60
615	test	200	endothall	145-73-3	O=C(O)C(C(C(=O)O)C(O1)CC2)C12	2.140	0.55	-1.11	-0.38	-0.06	1.89	-0.79	-0.46
616	test	201	N,N-dimethylaniline	121-69-7	N(c(cccc1)c1)(C)C	2.634	2.05	1.98	1.99	2.19	2.17	2.23	2.31
617	training	406	o-ethylaniline	578-54-1	Nc(c(ccc1)CC)c1	2.324	1.91	1.93	2.03	2.19	2.11	1.89	1.74
618	test	202	2,4,6-trimethylpyridine	108-75-8	n(c(cc(c1)C)C)c1C	2.400	2.10	2.06	1.73	1.53	2.45	1.84	1.88
619	test	203	benzeneethanamine	64-04-0	NCCc(cccc1)c1	2.144	1.41	1.16	1.26	1.92	1.34	1.30	1.41
620	training	407	1,5-cyclooctadiene	111-78-4	C(=CCCC=CC1)C1	3.096	3.58	3.50	2.76	2.65	3.73	3.52	3.16
621	training	408	vinylcyclohexene	100-40-3	C(=CCCC1C=C)C1	3.515	3.49	2.67	2.61	2.65	3.73	3.60	2.82
622	test	204	2,5-dimethyl-2,4-hexadiene	764-13-6	C(=CC=C(C)C)(C)C	3.281	3.55	3.47	3.11	3.04	3.95	3.13	3.56
623	training	409	butyl methacrylate	97-88-1	O=C(OCCCC)C(=C)C	2.944	2.59	2.49	2.49	1.95	2.75	2.03	2.88
624	test	205	diethyl succinate	123-25-1	O=C(OCC)CCC(=O)OCC	2.030	1.25	1.38	0.79	1.12	1.39	0.78	1.19
625	training	410	octanenitrile	124-12-9	C(#N)CCCCCCC	2.873	2.76	3.71	2.99	3.04	2.80	3.13	2.75
626	test	206	cyclooctane	292-64-8	C1CCCCCCC1	3.798	4.62	4.21	3.65	3.81	4.16	4.55	4.54
627	training	411	1-octene	111-66-0	C(=C)CCCCCC	3.863	4.61	3.87	3.63	3.15	4.13	4.53	4.57
628	test	207	2,4,4-trimethyl-1-pentene	107-39-1	C(=C)(CC(C)(C)C)C	3.852	4.03	3.81	3.17	3.15	4.08	3.99	3.68
629	training	412	2-octanone	111-13-7	O=C(CCCCCC)C	2.666	2.54	3.06	2.25	2.13	2.22	2.71	2.37

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
630	training	413	octanoic acid	124-07-2	O=C(O)CCCCCCC	3.036	2.92	2.81	2.72	2.97	3.03	3.02	3.05
631	test	208	1-bromooctane	111-83-1	BrCCCCCCCC	4.037	4.91	4.39	4.14	4.67	4.61	4.94	4.89
632	test	209	octane	111-65-9	C(CCCCCC)C	4.179	4.73	4.16	4.02	4.20	4.27	4.89	3.93
633	training	414	1-octanol	111-87-5	OCCCCCCCC	3.047	3.21	3.21	2.80	3.19	2.81	3.13	3.00
634	test	210	2-octanol	123-96-6	OC(CCCCCC)C	2.955	2.96	3.15	2.72	2.27	2.73	3.02	2.90
635	training	415	4-octanol	589-62-8	CCCCC(O)CCC	2.835	2.83	3.15	2.78	2.27	2.73	2.81	2.68
636	training	416	dibutyl ether	142-96-1	O(CCCC)CCCC	3.123	3.04	3.17	2.71	2.27	3.01	2.72	3.21
637	training	417	diethylene glycol diethyl ether	112-36-7	O(CCOCC)CCOCC	1.589	0.64	1.19	0.49	0.56	0.50	0.49	0.39
638	training	418	diethylene glycol monobutyl ether	112-34-5	O(CCOCCO)CCCC	1.682	0.63	1.23	0.71	0.56	0.29	0.47	0.56
639	training	419	octylamine	111-86-4	NCCCCCCCC	2.955	3.24	2.67	2.51	3.19	2.80	3.06	3.11
640	training	420	dibutylamine	111-92-2	N(CCCC)CCCC	2.917	2.71	2.70	2.44	2.27	2.77	2.52	2.83
641	training	421	octamethylcyclotetrasiloxane	556-67-2	C[Si]1(C)O[Si](C)(C)O[Si](C)(C)O[Si](C)(CO1	4.151	3.56	4.30	6.28	-0.64	5.09	3.89	4.02
642	test	211	folpet	133-07-3	O=C(N(SC(Cl)(Cl)Cl)C(=O)c1cccc2)c12	3.270	2.92	2.81	4.45	2.57	2.84	3.36	2.85
643	training	422	2H-1-benzopyran-2-one	91-64-5	c1cc2OC(=O)C=Cc2cc1	2.133	1.72	1.95	1.90	2.27	1.51	1.77	1.39
644	training	423	1H-indene-1,3(2H)-dione	606-23-5	O=C(c(c(C1=O)ccc2)c2)C1	1.709	1.54	1.79	1.33	1.72	0.75	0.98	0.61
645	test	212	isoquinoline	119-65-3	n(ccc(c1ccc2)c2)c1	2.509	2.14	2.09	1.59	2.07	2.14	2.04	2.08
646	training	424	quinoline	91-22-5	n(c(c(ccc1)cc2)c1)c2	2.481	2.19	2.29	2.02	2.07	2.14	2.12	2.03
647	training	425	cinnamonitrile	4360-47-8	C(#N)C=Cc(cccc1)c1	2.443	2.01	2.53	2.18	2.07	1.84	2.02	1.96
648	training	426	8-hydroxyquinoline	148-24-3	n(c(c(ccc1)cc2)c1O)c2	2.476	1.91	2.00	1.75	1.96	1.66	1.72	2.02
649	training	427	indene	95-13-6	c(c(C=C1)ccc2)(c2)C1	2.965	3.04	2.77	2.44	3.17	3.25	2.99	2.92
650	test	213	captan	133-06-2	O=C(N(SC(Cl)(Cl)Cl)C(=O)C1CC=CC2)C12	2.300	3.00	1.97	4.02	1.82	2.74	2.27	2.35
651	training	428	2-methylbenzofuran	4265-25-2	Cc2cc1ccccc1o2	3.129	3.07	2.79	2.28	2.15	3.09	2.30	2.73
652	training	429	2-propenophenone	768-03-6	C=CC(=O)c1ccccc1	2.400	1.82	2.07	2.27	2.33	2.03	2.15	1.88
653	test	214	cinnamic acid	621-82-9	O=C(O)C=Cc(cccc1)c1	2.536	2.38	1.63	1.90	2.00	2.07	1.91	2.13
654	training	430	(4-chloro-2-methylphenoxy)acetic acid	94-74-6	O=C(O)COc(c(cc(c1)Cl)C)c1	3.860	2.41	1.89	2.63	2.09	2.52	2.17	2.61
655	training	431	propanil	709-98-8	CCC(=O)Nc1ccc(Cl)c(Cl)c1	2.480	3.04	3.28	2.95	3.08	2.88	2.77	3.07
656	training	432	benzenepropanenitrile	645-59-0	N#CCCc(cccc1)c1	2.313	1.94	2.67	2.20	2.15	2.05	1.73	1.72
657	training	433	cinnamamide	621-79-4	O=C(N)C=Cc(cccc1)c1	2.144	1.19	1.09	1.30	1.59	0.82	1.18	1.43
658	training	434	indane	496-11-7	c(c(ccc1)CC2)(c1)C2	3.189	2.97	2.79	2.89	3.26	3.47	2.91	3.18
659	training	435	α-methylstyrene	98-83-9	c(C(=C)C)(cccc1)c1	3.270	3.31	2.68	2.83	3.17	3.44	3.31	3.48
660	training	436	chlorbromuron	13360-45-7	CON(C)C(=O)Nc1ccc(Br)c(Cl)c1	2.580	3.02	2.44	2.41	2.79	3.15	2.39	3.09
661	test	215	imidacloprid	105827-78-9	ClC1=CC=C(CN2CCN\C2=N\N(=O)=O)C=N1	2.640	0.65	1.11	1.77	1.62	0.56	2.43	1.15
662	training	437	diuron	330-54-1	O=C(N(C)C)Nc(ccc(c1Cl)Cl)c1	2.820	2.92	2.82	2.48	2.65	2.67	2.30	2.68
663	training	438	linuron	330-55-2	O=C(N(OC)C)Nc(ccc(c1Cl)Cl)c1	2.430	2.82	2.36	2.33	2.65	2.91	2.21	3.20
664	training	439	2,3-dihydro-1H-inden-1-ol	6351-10-6	c(ccc1C2O)cc1CC2	4.060	1.59	2.02	1.78	2.24	1.93	1.52	1.46
665	test	216	2,3-dihydro-1H-inden-5-ol	1470-94-6	Oc(ccc(c1CC2)C2)c1	4.060	2.37	2.49	2.62	2.51	2.99	2.51	2.41
666	test	217	4-methylacetophenone	122-00-9	O=C(c(ccc(c1)C)c1)C	2.568	2.11	2.22	2.06	2.42	2.22	2.30	2.10
667	training	440	1-phenyl-1-propanone	93-55-0	O=C(c(cccc1)c1)CC	2.568	2.15	2.37	2.24	2.42	2.16	2.12	2.19
668	training	441	1-phenyl-2-propanone	103-79-7	O=C(Cc(cccc1)c1)C	2.160	1.70	2.02	1.46	2.15	1.47	1.71	1.44
669	training	442	ethyl benzoate	93-89-0	O=C(OCC)c(cccc1)c1	2.813	2.39	2.39	2.03	2.35	2.32	2.38	2.64
670	test	218	benzyl acetate	140-11-4	O=C(OCc(cccc1)c1)C	2.443	2.07	1.85	1.61	2.08	2.08	1.95	1.96

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
671	training	443	4-methylphenyl acetate	140-39-6	O=C(Oc(ccc(c1)C)c1)C	2.525	1.96	2.23	2.08	2.35	2.14	2.24	2.11
672	training	444	(±)-2-phenylpropionic acid	492-37-5	O=C(O)C(c(cccc1)c1)C	2.356	2.17	1.72	1.93	2.08	1.85	1.66	1.91
673	training	445	ethyl vanillin	121-32-4	O=Cc(ccc(O)c1OCC)c1	2.237	1.82	1.70	1.65	1.24	1.55	1.65	1.58
674	test	219	metobromuron	3060-89-7	CON(C)C(=O)Nc1ccc(Br)cc1	2.020	2.18	1.83	1.75	2.24	2.51	1.77	2.38
675	training	446	monuron	150-68-5	O=C(N(C)C)Nc(ccc(c1)Cl)c1	1.700	1.96	2.21	1.82	2.09	2.03	1.68	1.94
676	training	447	3-(4-chlorophenyl)-1-methoxy-1-methylurea	1746-81-2	O=C(N(OC)C)Nc(ccc(c1)Cl)c1	1.840	1.99	1.74	1.66	2.09	2.26	1.59	2.30
677	training	448	chlorpyrifos	2921-88-2	CCOP(=S)(OCC)Oc1nc(Cl)c(Cl)cc1Cl	3.790	5.15	5.11	4.60	3.21	4.66	5.45	5.27
678	training	449	1,1-dimethyl-3-(3-fluorophenyl) urea	330-39-2	CN(C)C(=O)Nc1cccc(F)c1	1.730	1.32	1.66	1.36	1.94	1.58	1.22	1.37
679	training	450	1,1-dimethyl-3-(4-fluorophenyl) urea	332-33-2	CN(C)C(=O)Nc1ccc(F)cc1	1.430	1.46	1.66	1.36	1.94	1.58	1.22	1.13
680	training	451	1,2,3,4-tetrahydroquinoline	635-46-1	N(c(c(ccc1)CC2)c1)C2	2.623	2.27	2.16	2.17	2.12	2.55	1.87	2.29
681	test	220	p-dimethylaminobenzaldehyde	100-10-7	O=Cc(ccc(N(C)C)c1)c1	2.362	1.80	1.67	1.75	1.81	1.89	1.93	1.81
682	training	452	cumene	98-82-8	c(cccc1)(c1)C(C)C	3.368	3.67	3.08	3.02	3.26	3.45	3.43	3.66
683	training	453	m-ethyltoluene	620-14-4	CCc1cccc(C)c1	3.542	3.79	2.97	3.26	3.26	3.58	3.36	3.63
684	training	454	o-ethyltoluene	611-14-3	c(c(ccc1)C)(c1)CC	3.297	3.87	2.97	3.26	3.26	3.58	3.36	3.53
685	training	455	p-ethyltoluene	622-96-8	c(ccc(c1)C)(c1)CC	3.352	3.83	2.97	3.26	3.26	3.58	3.36	3.58
686	test	221	1,2,3-trimethylbenzene	526-73-8	c(c(c(cc1)C)C)(c1)C	3.335	3.63	2.93	3.29	3.26	3.63	3.33	3.62
687	training	456	1,2,4-trimethylbenzene	95-63-6	c(ccc(c1C)C)(c1)C	3.352	3.62	2.93	3.29	3.26	3.63	3.33	3.02
688	training	457	mesitylene	108-67-8	c(cc(cc1C)C)(c1)C	3.237	3.64	2.93	3.29	3.26	3.63	3.33	3.42
689	test	222	propylbenzene	103-65-1	c(cccc1)(c1)CCC	3.384	3.86	3.12	3.23	3.26	3.52	3.49	3.72
690	training	458	fenitrothion	122-14-5	COP(=S)(OC)Oc1ccc(N(=O)=O)c(C)c1	3.510	3.31	3.09	2.69	1.92	3.30	3.30	3.30
691	training	459	benzyl ethyl ether	539-30-0	O(Cc(cccc1)c1)CC	2.552	2.38	2.25	1.98	2.24	2.27	2.15	2.29
692	training	460	benzenepropanol	122-97-4	OCCCc(cccc1)c1	2.400	2.00	2.16	2.00	2.24	2.06	1.73	1.88
693	test	223	2-propylphenol	644-35-9	Oc(c(ccc1)CCC)c1	2.971	2.95	2.82	2.96	2.51	3.04	2.87	2.93
694	training	461	4-propylphenol	645-56-7	Oc(ccc(c1)CCC)c1	3.118	3.01	2.82	2.96	2.51	3.04	3.09	3.20
695	test	224	2,3,4-trimethylphenol	526-85-2	Oc(c(c(c(c1)C)C)C)c1	3.760	2.75	2.63	3.02	2.51	3.15	2.72	2.66
696	training	462	2,3,5-trimethylphenol	697-82-5	Oc(c(c(cc1C)C)C)c1	3.760	2.73	2.63	3.02	2.51	3.15	2.72	2.66
697	training	463	2,3,6-trimethylphenol	2416-94-6	Oc(c(ccc1C)C)c1C	3.760	2.72	2.63	3.02	2.51	3.15	2.50	2.66
698	test	225	2,4,5-trimethylphenol	496-78-6	Oc(c(cc(c1C)C)C)c1	3.760	2.75	2.63	3.02	2.51	3.15	2.72	2.66
699	training	464	2,4,6-trimethylphenol	527-60-6	Oc(c(cc(c1)C)C)c1C	3.760	2.72	2.63	3.02	2.51	3.15	2.50	2.66
700	training	465	3,4,5-trimethylphenol	527-54-8	Oc(cc(c(c1C)C)C)c1	3.760	2.77	2.63	3.02	2.51	3.15	2.93	2.97
701	training	466	bromacil	314-40-9	N1C(=O)N(C(C)CC)C(=O)C(Br)=C1C	1.970	1.20	2.06	1.41	1.80	1.68	1.46	2.11
702	training	467	terbacil	5902-51-2	CC1=C(Cl)C(=O)N(C(C)(C)C)C(=O)N1	1.630	1.78	1.70	1.01	1.65	1.75	1.34	1.89
703	test	226	cyanazine	21725-46-2	CCNc1nc(Cl)nc(NC(C)(C)C(#N))n1	2.260	2.05	2.18	2.34	2.03	2.51	0.91	2.22
704	training	468	amphetamine	300-62-9	NC(C)Cc(cccc1)c1	2.334	1.85	1.57	1.63	2.24	1.76	1.76	1.76
705	training	469	N,N-dimethylbenzylamine	103-83-3	N(Cc(cccc1)c1)(C)C	2.454	1.84	1.56	1.90	2.24	1.75	1.77	1.98
706	training	470	isophorone	78-59-1	O=C(C=C(CC1(C)C)C)C1	2.302	1.90	2.50	2.06	1.95	2.62	1.95	1.64
707	test	227	glyceryl triacetate	102-76-1	O=C(OCC(OC(=O)C)COC(=O)C)C	1.513	0.40	0.39	-0.27	0.59	0.36	0.34	0.25
708	test	228	propazine	139-40-2	CC(C)Nc1nc(Cl)nc(NC(C)C)n1	2.190	2.94	2.88	2.91	2.89	3.24	2.12	2.93
709	test	229	triethazine	1912-26-1	n(c(nc(n1)N(CC)CC)NCC)c1Cl	2.740	3.58	2.94	2.86	2.89	3.44	2.19	3.34
710	test	230	azelaic acid	123-99-9	O=C(O)CCCCCCCC(=O)O	2.231	1.37	1.92	1.87	1.44	1.70	1.71	1.57
711	training	471	ametryn	834-12-8	CCNc1nc(NC(C)C)nc(SC)n1	2.130	3.09	2.36	2.74	2.59	3.32	1.78	2.98
712	training	472	1-nonene	124-11-8	C(=C)CCCCCCC	4.179	5.14	4.33	4.09	4.38	4.62	5.10	5.15

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
713	test	231	2-nonanone	821-55-6	O=C(CCCCCCC)C	3.096	3.08	3.53	2.70	3.36	2.71	3.28	3.14
714	training	473	5-methyl-2-octanone	58654-67-4	O=C(CCC(CCC)C)C	2.965	3.07	3.40	2.50	2.45	2.64	3.22	2.92
715	test	232	nonanoic acid	112-05-0	O=C(O)CCCCCCCC	3.237	3.47	3.27	3.17	3.29	3.52	3.59	3.55
716	test	233	nonane	111-84-2	C(CCCCCCC)C	4.451	5.24	4.63	4.48	4.52	4.76	5.46	4.48
717	training	474	1-nonanol	143-08-8	OCCCCCCCCC	3.564	3.76	3.67	3.25	3.50	3.30	3.70	4.26
718	test	234	2,6-dimethyl-4-heptanol	108-82-7	OC(CC(C)C)CC(C)C	3.053	3.03	3.36	2.83	2.59	3.08	3.25	2.98
719	training	475	tripropylamine	102-69-2	N(CCC)(CCC)CCC	2.895	3.08	2.89	2.94	2.59	2.99	2.41	2.79
720	training	476	terbufos	13071-79-9	CCOP(=S)(OCC)SCSC(C)(C)C	2.500	4.61	4.85	3.52	1.97	4.24	4.50	4.48
721	training	477	ethion	563-12-2	CCOP(=S)(OCC)SCSP(=S)(OCC)OCC	3.940	4.74	4.63	3.78	0.63	4.75	4.98	5.07
722	training	478	chlordane	57-74-9	ClC1CC2C(C1Cl)C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	5.150	6.02	5.77	4.78	5.47	6.26	4.36	4.94
724	training	479	a-chlordan	5103-71-9	ClC1CC2C(C1Cl)C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	5.150	6.02	5.77	4.78	5.47	6.26	4.36	4.94
725	test	235	trans-chlordane	5103-74-2	ClC1CC2C(C1Cl)C3(Cl)C(=C(Cl)C2(Cl)C3(Cl)Cl)Cl	5.150	6.02	5.77	4.78	5.47	6.26	4.36	4.94
726	training	480	2-hydroxy-1,4-naphthalenedione	83-72-7	c1ccc2C(=O)C=C(O)C(=O)c2c1	2.171	0.99	1.41	1.20	1.37	0.78	1.79	1.38
727	test	236	1-chloronaphthalene	90-13-1	c(c(c(cc1)Cl)ccc2)(c2)c1	3.499	3.95	3.78	3.40	3.94	3.81	3.91	4.24
728	test	237	2-chloronaphthalene	91-58-7	c(c(ccc1Cl)ccc2)(c2)c1	3.542	3.91	3.78	3.40	3.94	3.81	3.91	4.14
729	training	481	naphthalene	91-20-3	c(c(ccc1)ccc2)(c1)c2	3.000	3.33	3.16	2.74	3.39	3.17	3.29	3.30
730	test	238	azulene	275-51-4	C1=CC=C2C=CC=C2C=C1	3.129	3.51	3.08	2.74	2.99	3.38	2.17	3.20
731	training	482	1-naphthol	90-15-3	Oc(c(c(ccc1)cc2)c1)c2	2.922	2.79	2.86	2.47	2.64	2.69	2.88	2.84
732	test	239	2-naphthol	135-19-3	Oc(ccc(c1ccc2)c2)c1	2.846	2.93	2.86	2.47	2.64	2.69	2.88	2.70
733	training	483	captafol	2425-06-1	ClC(Cl)C(Cl)(Cl)SN1C(=O)C2CC=CCC2C1=O	3.320	3.57	2.14	3.51	2.36	3.42	2.77	3.83
734	test	240	2-methylquinoline	91-63-4	n(c(c(ccc1)cc2)c1)c2C	2.786	2.66	2.71	2.30	2.37	2.69	2.44	2.59
735	test	241	1-naphthylamine	134-32-7	c(c(c(N)cc1)ccc2)(c2)c1	3.580	2.27	2.44	1.99	2.64	2.25	2.47	2.25
736	training	484	2-naphthylamine	91-59-8	c(c(ccc1N)ccc2)(c2)c1	3.580	2.30	2.44	1.99	2.64	2.25	2.47	2.28
737	test	242	metamitron	41394-05-2	c1ccccc1C2=NN=C(C)N(N)C2(=O)	1.550	1.17	0.72	0.53	1.96	1.44	1.54	0.72
738	training	485	benzalacetone	122-57-6	O=C(C=Cc(cccc1)c1)C	2.503	2.23	2.35	1.89	2.37	2.04	2.13	2.07
739	training	486	methyl cinnamate	1754-62-7	c(ccc1C=CC(=O)OC)cc1	2.802	2.58	2.08	2.15	2.30	2.36	2.23	2.62
740	test	243	dimethyl phthalate	131-11-3	O=C(OC)c(c(ccc1)C(=O)OC)c1	1.630	1.96	1.93	1.54	2.01	1.66	1.89	1.56
741	test	244	dimethyl terephthalate	120-61-6	O=C(OC)c(ccc(c1)C(=O)OC)c1	2.601	1.83	1.93	1.54	2.01	1.66	1.89	2.25
742	training	487	Fluometuron	2164-17-2	O=C(N(C)C)Nc(cccc1C(F)(F)F)c1	1.960	2.16	2.36	2.10	2.52	2.35	1.98	2.42
743	training	488	1,2,3,4-tetrahydronaphthalene	119-64-2	c(c(ccc1)CCC2)(c1)C2	3.276	3.79	3.11	3.34	3.56	3.96	3.48	3.49
744	training	489	3-phenyl-1-cyclopropyl urea	13140-86-8	c1ccccc1NC(=O)NC2CC2	1.720	1.62	2.01	1.44	1.43	1.97	1.46	1.65
745	training	490	azinphos-methyl	86-50-0	S=P(OC)(OC)SCN1N=Nc2ccccc2C1(=O)	2.690	2.75	1.86	2.93	0.91	2.53	2.22	2.75
746	training	491	isopropyl benzoate	939-48-0	O=C(OC(C)C)c(cccc1)c1	3.107	2.72	2.79	2.41	2.65	2.74	2.84	3.18
747	test	245	Chlorotoluron	15545-48-9	CN(C)C(=O)Nc1ccc(C)c(Cl)c1	2.430	2.25	2.53	2.30	2.38	2.58	1.91	2.41
749	training	492	isopropyl phenylcarbamate	122-42-9	O=C(OC(C)C)Nc(cccc1)c1	1.950	2.60	2.69	2.32	1.82	2.66	2.33	2.60
750	training	493	butylbenzene	104-51-8	c(cccc1)(c1)CCCC	3.694	4.34	3.58	3.68	3.56	4.01	4.06	4.38
751	training	494	isobutylbenzene	538-93-2	c(cccc1)(c1)CC(C)C	3.558	4.13	3.46	3.48	3.56	3.94	4.00	3.94
752	test	246	sec-butylbenzene	135-98-8	c(cccc1)(c1)C(CC)C	3.863	4.36	3.54	3.48	3.56	3.94	4.00	4.20
753	training	495	tert-butylbenzene	98-06-6	c(cccc1)(c1)C(C)(C)C	3.613	4.49	3.50	3.23	3.56	3.90	3.84	4.11

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
754	training	496	m-cymene	535-77-3	c(cccc1C(C)C)(c1)C	3.825	4.07	3.39	3.51	3.56	4.00	3.86	3.96
755	test	247	o-cymene	527-84-4	c(c(ccc1)C)(c1)C(C)C	3.760	4.11	3.39	3.51	3.56	4.00	3.86	3.41
756	training	497	p-cymene	99-87-6	c(ccc(c1)C)(c1)C(C)C	3.607	4.17	3.39	3.51	3.56	4.00	3.86	4.10
757	training	498	o-diethylbenzene	135-01-3	CCc1ccccc1CC	3.781	4.55	3.33	3.71	3.56	4.07	3.82	3.72
758	training	499	m-diethylbenzene	141-93-5	c(cccc1CC)(c1)CC	3.863	4.38	3.33	3.71	3.56	4.07	3.82	3.95
759	test	248	p-diethylbenzene	105-05-5	c(ccc(c1)CC)(c1)CC	3.869	4.36	3.33	3.71	3.56	4.07	3.82	3.52
760	test	249	3-ethyl-o-xylene	933-98-2	c(c(c(cc1)C)C)(c1)CC	3.738	4.40	3.28	3.74	3.56	4.13	3.80	3.89
761	test	250	4-ethyl-o-xylene	934-80-5	c(ccc(c1C)C)(c1)CC	3.825	4.33	3.28	3.74	3.56	4.13	3.80	3.45
762	training	500	2-ethyl-m-xylene	2870-04-4	c(c(c(cc1)C)CC)(c1)C	3.705	4.40	3.28	3.74	3.56	4.13	3.80	3.45
763	training	501	4-ethyl-m-xylene	874-41-9	c(ccc(c1C)CC)(c1)C	3.809	4.32	3.28	3.74	3.56	4.13	3.80	3.89
764	training	502	5-ethyl-m-xylene	934-74-7	c(cc(cc1C)CC)(c1)C	3.852	4.34	3.28	3.74	3.56	4.13	3.80	3.45
765	test	251	2-ethyl-p-xylene	1758-88-9	c(ccc(c1CC)C)(c1)C	3.787	4.35	3.28	3.74	3.56	4.13	3.80	3.45
766	training	503	1,2,3,4-tetramethylbenzene	488-23-3	c(c(c(c(c1)C)C)C)(c1)C	3.553	4.07	3.24	3.78	3.56	4.18	3.77	3.98
767	training	504	1,2,3,5-tetramethylbenzene	527-53-7	c(cc(c(c1C)C)C)(c1)C	3.607	4.06	3.24	3.78	3.56	4.18	3.77	4.04
768	test	252	1,2,4,5-tetramethylbenzene	95-93-2	c(c(cc(c1C)C)C)(c1)C	3.607	4.05	3.24	3.78	3.56	4.18	3.77	4.00
769	training	505	parathion	56-38-2	CCOP(=S)(OCC)Oc1ccc(cc1)N(=O)=O	3.020	3.76	3.65	2.90	2.22	3.73	3.92	3.83
770	training	506	1,1-dimethyl-3-(4-methoxy-phenyl) urea	28170-54-9	COc1cccc(NC(=O)N(C)C)c1	1.720	1.18	1.49	1.14	1.28	1.46	0.97	1.29
771	test	253	4-butylphenol	1638-22-8	Oc(ccc(c1)CCCC)c1	3.363	3.55	3.28	3.42	2.81	3.53	3.65	3.38
772	test	254	N,N-dietyhlaniline	91-66-7	N(c(cccc1)c1)(CC)CC	3.178	3.43	2.85	2.69	2.81	3.15	3.08	3.31
773	training	507	α-pinene	80-56-8	C(C(CC1C2)C1(C)C)(=C2)C	4.005	3.66	2.99	2.87	3.37	4.27	3.12	2.85
774	training	508	γ-terpinene	99-85-4	C(=CCC(=C1)C)(C1)C(C)C	3.825	4.36	3.79	3.45	3.27	4.75	2.80	2.76
775	test	255	terpinolene	586-62-9	C(=C(C)C)(CCC(=C1)C)C1	3.809	3.82	3.80	3.64	3.27	4.88	2.27	2.85
776	training	509	ipazine	1912-25-0	CCN(CC)c1nc(Cl)nc(NC(C)C)n1	3.390	3.65	3.34	3.24	3.18	3.86	2.65	3.77
777	training	510	methoxypropazine	1610-18-0	O(c(nc(nc1NC(C)C)NC(C)C)n1)C	2.430	2.80	2.35	2.56	2.53	3.57	2.06	2.99
778	training	511	prometryn	7287-19-6	CSc1nc(NC(C)C)nc(NC(C)C)n1	2.800	3.31	2.77	3.12	2.89	3.73	2.24	3.51
779	training	512	2-decanone	693-54-9	O=C(CCCCCCCC)C	3.428	3.63	3.99	3.16	3.66	3.20	3.85	3.73
780	training	513	decanoic acid	334-48-5	CCCCCCCCCC(=O)O	3.602	3.93	3.74	3.63	3.59	4.02	4.16	4.09
781	test	256	decane	124-18-5	C(CCCCCCCC)C	4.777	5.87	5.09	4.93	4.82	5.25	4.97	5.02
782	training	514	2,2,3,3-tetramethylhexane	13475-81-5	CCCC(C)(C)C(C)(C)C	4.113	5.64	4.95	4.03	4.82	5.03	4.93	4.76
783	training	515	1-decanol	112-30-1	OCCCCCCCCCC	3.863	4.24	4.14	3.71	3.81	3.79	4.26	4.57
784	test	257	2-(2-furyl)benzimidazole	3878-19-1	n1c2ccccc2nc1c3ccco3	2.550	2.84	2.36	2.67	1.82	2.37	1.97	2.25
785	training	516	2-methyl-1,4-naphthalenedione	58-27-5	c1cc2C(=O)C=C(C)C(=O)c2cc1	2.574	1.91	2.40	2.20	2.23	2.21	1.58	2.20
786	training	517	2-hydroxy-3-methyl-1,4-naphthalenedione	483-55-6	CC2=C(O)C(=O)c1ccccc1C2=O	2.030	1.48	1.81	1.64	1.66	1.32	1.93	1.20
787	test	258	2-methoxy-1,4-naphthalenedione	2348-82-5	c1ccc2C(=O)C=C(OC)C(=O)c2c1	2.111	1.70	1.87	1.26	1.66	0.95	1.68	1.35
788	training	518	4-phenylpyridine	939-23-1	n(ccc(c1)c(cccc2)c2)c1	2.786	2.40	2.59	2.20	2.19	2.57	2.71	2.59
789	training	519	1-methylnaphthalene	90-12-0	c(c(c(cc1)C)ccc2)(c2)c1	3.482	3.84	3.48	3.22	3.68	3.72	3.72	3.87
790	test	259	2-methylnaphthalene	91-57-6	c(c(ccc1C)ccc2)(c2)c1	3.553	3.83	3.48	3.22	3.68	3.72	3.72	3.86
791	test	260	1-naphthalenemethanol	4780-79-4	c1ccc2ccccc2c1CO	2.220	2.17	2.54	2.13	2.66	2.25	2.48	2.39
792	training	520	2-naphthalenemethanol	1592-38-7	OCc1cc2ccccc2cc1	2.220	2.28	2.54	2.13	2.66	2.25	2.48	2.21
793	training	521	chloramphenicol	56-75-7	O=C(NC(C(O)c(ccc(N(=O)(=O))c1)c1)CO)C(Cl)Cl	1.997	1.15	0.67	1.02	1.23	0.92	0.69	1.14
794	training	522	butyl benzoate	136-60-7	O=C(OCCCC)c(cccc1)c1	3.466	3.40	3.32	3.01	2.95	3.30	3.31	3.84
795	test	261	methiocarb	2032-65-7	CNC(=O)Oc1cc(C)c(SC)c(C)c1	2.250	2.54	2.83	3.11	2.38	2.87	2.64	2.92

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
796	training	523	pentylbenzene	538-68-1	c(cccc1)(c1)CCCCC	4.043	4.81	4.05	4.14	3.85	4.50	4.63	4.98
797	training	524	pentamethylbenzene	700-12-9	c(c(c(c(c1C)C)C)C)(c1)C	3.858	4.42	3.56	4.26	3.85	4.73	4.21	4.56
798	test	262	2-undecanone	112-12-9	O=C(CCCCCCCCC)C	3.602	4.25	4.45	3.62	3.95	3.69	4.42	4.09
799	test	263	methyl decanoate	110-42-9	O=C(OC)CCCCCCCCC	3.776	4.58	4.19	3.88	3.88	4.30	4.48	4.67
800	test	264	1-undecanol	112-42-5	OCCCCCCCCCCC	3.945	4.83	4.60	4.16	4.10	4.28	4.83	4.59
801	test	265	decachlorobiphenyl	2051-24-3	Clc1c(Cl)c(Cl)c(c(Cl)c1Cl)c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl	5.870	8.59	9.79	9.99	7.37	10.20	10.18	8.27
802	training	525	2,2′,3,3′,4,5,5′,6,6′nonachlorobiphenyl	52663-77-1	Clc1cc(Cl)c(Cl)c(c1Cl)c2c(Cl)c(Cl)c(Cl)c(Cl)c2Cl	5.816	8.34	9.18	9.33	7.15	9.56	9.56	8.16
803	training	526	2,2′,3,3′,5,5′,6,6′-octachlorobiphenyl	2136-99-4	Clc1cc(Cl)c(Cl)c(c1Cl)c2c(Cl)c(Cl)cc(Cl)c2Cl	5.239	8.07	8.57	8.66	6.93	8.91	8.94	7.73
804	training	527	2,2′,3,3′,4,4′,6-heptachlorobiphenyl	52663-71-5	Clc1ccc(c(Cl)c1Cl)c2c(Cl)cc(Cl)c(Cl)c2Cl	5.022	7.69	7.95	8.00	6.70	8.27	8.32	7.45
805	test	266	fiponil	120068-37-3	Clc1cc(C(F)(F)F)cc(Cl)c1N2C(N)=C(S(=O)C(F)(F)F)C(C#N)=N2	3.080	4.40	4.36	5.58	2.65	6.64	3.71	4.53
806	training	528	2,2′,3,3′,4,4′hexachlorobiphenyl	38380-07-3	Clc1ccc(c(Cl)c1Cl)c2ccc(Cl)c(Cl)c2Cl	3.830	7.27	7.34	7.34	6.47	7.62	7.69	7.32
807	test	267	2,2′,4,4′,6,6′-hexachlorobiphenyl	33979-03-2	Clc1cc(Cl)c(c(Cl)c1)c2c(Cl)cc(Cl)cc2Cl	3.830	7.27	7.34	7.34	6.47	7.62	7.69	7.29
808	training	529	2,2′,3,3′,6,6′-hexachlorobiphenyl	38411-22-2	Clc1ccc(Cl)c(c1Cl)c2c(Cl)ccc(Cl)c2Cl	3.830	7.28	7.34	7.34	6.47	7.62	7.69	7.12
809	training	530	2,3,3′,4,4′,5-hexachlorobiphenyl	38380-08-4	Clc1c(c(cc(c1Cl)c2cc(c(cc2)Cl)Cl)Cl)Cl	3.830	7.26	7.34	7.34	6.47	7.62	7.69	7.57
810	test	268	2,2′,4,4′,5′5′-hexachloro-1,1′-biphenyl	35065-27-1	Clc1cc(Cl)c(cc1Cl)c2cc(Cl)c(Cl)cc2Cl	3.830	7.25	7.34	7.34	6.47	7.62	7.69	7.16
811	training	531	3,3′,4,4′,5,5′-hexachlorobiphenyl	32774-16-6	Clc1cc(cc(Cl)c1Cl)c2cc(Cl)c(Cl)c(Cl)c2	3.830	7.26	7.34	7.34	6.47	7.62	7.69	7.41
812	training	532	2,3,4,5,6-pentachlorobiphenyl	18259-05-7	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1c2ccccc2	4.804	6.79	6.73	6.67	6.23	6.98	7.07	6.74
813	training	533	2,2′,4,5,5′-pentachlorobiphenyl	37680-73-2	Clc1ccc(Cl)c(c1)c2cc(Cl)c(Cl)cc2Cl	4.859	6.77	6.73	6.67	6.23	6.98	7.07	6.50
815	training	534	2,3,4,5-tetrachlorobiphenyl	33284-53-6	Clc2cc(c1ccccc1)c(Cl)c(Cl)c2Cl	5.640	6.14	6.12	6.01	5.99	6.34	6.45	6.41
816	training	535	2,2′,4′,5-tetrachlorobiphenyl	41464-40-8	Clc1ccc(c(Cl)c1)c2cc(Cl)ccc2Cl	5.640	6.23	6.12	6.01	5.99	6.34	6.45	6.36
818	training	536	3,3′,4,4′-tetrachlorobiphenyl	32598-13-3	Clc1ccc(cc1Cl)c2ccc(Cl)c(Cl)c2	5.000	6.21	6.12	6.01	5.99	6.34	6.45	6.63
819	test	269	2,2′,3,3′-tetrachlorobiphenyl	38444-93-8	Clc1cccc(c1Cl)c2cccc(Cl)c2Cl	5.000	6.21	6.12	6.01	5.99	6.34	6.45	6.18
820	training	537	2,2′,5,5′-tetrachlorobiphenyl	35693-99-3	c1c(Cl)ccc(Cl)c1c2c(Cl)ccc(Cl)c2	5.370	6.24	6.12	6.01	5.99	6.34	6.45	6.09
821	test	270	2,2′,6,6′-tetrachlorobiphenyl	15968-05-5	c(c(c(cc1)Cl)c(c(ccc2)Cl)c2Cl)(c1)Cl	4.890	6.22	6.12	6.01	5.99	6.34	6.45	5.94
822	test	271	2,3′,4′,5-tetrachlorobiphenyl	32598-11-1	Clc1ccc(Cl)c(c1)c2ccc(Cl)c(Cl)c2	4.850	6.22	6.12	6.01	5.99	6.34	6.45	6.23
823	training	538	2,4,5-trichlorobiphenyl	15862-07-4	Clc1cc(Cl)c(cc1Cl)c2ccccc2	5.210	5.69	5.50	5.34	5.47	5.69	5.83	5.90
824	test	272	2,4,6-trichlorobiphenyl	35693-92-6	Clc1cc(Cl)c(c(Cl)c1)c2ccccc2	5.210	5.70	5.50	5.34	5.47	5.69	5.83	5.71
825	training	539	2,2′,5-trichlorobiphenyl	37680-65-2	Clc1ccc(Cl)c(c1)c2ccccc2Cl	5.210	5.71	5.50	5.34	5.47	5.69	5.83	5.60
826	training	540	2,3′,4′-trichlorobiphenyl	38444-86-9	c1c(Cl)c(Cl)ccc1c2c(Cl)cccc2	5.210	5.69	5.50	5.34	5.47	5.69	5.83	5.87
827	test	273	2,3,5-trichlorobiphenyl	38444-81-4	Clc1cccc(c1)c2cc(Cl)ccc2Cl	5.210	5.70	5.50	5.34	5.47	5.69	5.83	5.76
828	training	541	2,4,4′-trichlorobiphenyl	7012-37-5	Clc1cc(Cl)ccc1c2ccc(Cl)cc2	5.210	5.70	5.50	5.34	5.47	5.69	5.83	5.62
829	training	542	2,4′,5-trichlorobiphenyl	16606-02-3	Clc1ccc(cc1)c2cc(Cl)ccc2Cl	5.210	5.70	5.50	5.34	5.47	5.69	5.83	5.79
831	test	274	2,5-dichlorobiphenyl	34883-39-1	Clc1ccc(Cl)c(c1)c2ccccc2	4.700	5.14	4.89	4.68	4.95	5.05	5.21	5.16
832	training	543	2,6-dichlorobiphenyl	33146-45-1	Clc1cccc(Cl)c1c2ccccc2	4.700	5.15	4.89	4.68	4.95	5.05	5.21	4.98
833	training	544	3,3′-dichloro-1,1′-biphenyl	2050-67-1	c1ccc(Cl)cc1c2cc(Cl)ccc2	4.700	5.13	4.89	4.68	4.95	5.05	5.21	5.21
834	test	275	4,4′-dichloro-1,1′-biphenyl	2050-68-2	Clc1ccc(cc1)c2ccc(Cl)cc2	5.270	5.12	4.89	4.68	4.95	5.05	5.21	5.58
835	test	276	2,2′-dichloro-1,1′-biphenyl	13029-08-8	Clc1ccccc1c2ccccc2Cl	4.700	5.15	4.89	4.68	4.95	5.05	5.21	4.97
836	training	545	3,4-dichloro-1,1′-biphenyl	2974-92-7	c1ccccc1c2cc(Cl)c(Cl)cc2	4.700	5.13	4.89	4.68	4.95	5.05	5.21	5.29
838	training	546	2,4′-dichloro-1,1′-biphenyl	34883-43-7	Clc1ccc(cc1)c2ccccc2Cl	4.550	5.14	4.89	4.68	4.95	5.05	5.21	5.10

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
839	training	547	dieldrin	60-57-1	ClC4=C(Cl)C5(Cl)C3C1CC(C2OC12)C3C4(Cl)C5(Cl)Cl	4.110	4.98	4.21	3.40	4.45	5.45	3.01	3.67
840	test	277	endrin	72-20-8	ClC4=C(Cl)C5(Cl)C3C1CC(C2OC12)C3C4(Cl)C5(Cl)Cl	4.200	4.98	4.21	3.40	4.45	5.45	3.01	3.67
841	test	278	1,10-phenanthroline	66-71-7	n(c(c(ccc1cccn2)cc3)c12)c3	2.373	2.31	2.61	2.20	1.90	2.29	2.23	1.78
842	test	279	phenazine	92-82-0	n(c(c(nc1cccc2)ccc3)c3)c12	2.922	2.82	3.02	3.06	1.90	2.29	2.89	2.84
843	test	280	dibenzofuran	132-64-9	o(c(c(c1cccc2)ccc3)c3)c12	3.618	3.92	3.76	3.33	2.86	3.71	3.42	4.12
844	training	548	2-chlorobiphenyl	2051-60-7	Clc1ccccc1c2ccccc2	3.836	4.59	4.28	4.01	4.42	4.40	4.58	4.53
845	training	549	3-chlorobiphenyl	2051-61-8	Clc1cccc(c1)c2ccccc2	3.869	4.57	4.28	4.01	4.42	4.40	4.58	4.58
846	training	550	4-chlorobiphenyl	2051-62-9	Clc1ccc(cc1)c2ccccc2	3.885	4.57	4.28	4.01	4.42	4.40	4.58	4.61
847	test	281	norflurazon	27314-13-2	c1c(C(F)(F)F)cccc1N2C(=O)C(Cl)=C(NC)C=N2	2.660	2.66	2.96	2.02	2.90	2.19	3.20	2.30
848	training	551	dibenzopyrrole	86-74-8	n(c(c(c1cccc2)ccc3)c3)c12	3.401	3.69	3.44	3.32	2.86	3.23	3.53	3.72
849	test	282	acenaphthene	83-32-9	c(c(ccc1)ccc2)(c1CC3)c23	3.531	4.01	3.66	3.34	3.96	4.15	3.61	3.92
850	test	283	phenylbenzene	92-52-4	c(c(cccc1)c1)(cccc2)c2	3.040	4.02	3.66	3.35	3.88	3.76	3.96	4.01
852	training	552	azobenzene	17082-12-1	c(ccc1N=Nc(ccc2)cc2)cc1	3.130	4.30	4.34	4.20	3.25	4.11	4.14	3.82
853	training	553	diphenyl ether	101-84-8	O(c(cccc1)c1)c(cccc2)c2	3.667	3.68	3.39	3.39	3.39	4.05	3.50	4.21
854	training	554	diphenyl sulfide	139-66-2	S(c(cccc1)c1)c(cccc2)c2	3.798	4.36	4.07	3.95	4.41	4.29	4.22	4.45
855	test	284	p-aminodiphenyl	92-67-1	Nc(ccc(c(cccc1)c1)c2)c2	2.933	2.89	2.94	2.60	3.13	2.84	3.14	2.86
856	test	285	diphenylamine	122-39-4	N(c(cccc1)c1)c(cccc2)c2	2.780	3.34	3.86	3.38	3.39	3.29	3.23	3.50
857	training	555	carbaryl	63-25-2	O=C(Oc(c(c(ccc1)cc2)c1)c2)NC	2.020	2.45	2.88	2.50	2.22	2.35	2.71	2.36
858	training	556	p-aminoazobenzene	60-09-3	N(=Nc(cccc1)c1)c(ccc(N)c2)c2	3.232	4.02	3.62	3.45	2.68	3.19	3.32	3.41
859	training	557	1,2-dimethylnaphthalene	573-98-8	Cc2ccc1ccccc1c2C	3.722	4.38	3.79	3.71	3.96	4.26	4.16	4.31
860	training	558	1,3-dimethylnaphthalene	575-41-7	Cc2cc(C)c1ccccc1c2	3.781	4.36	3.79	3.71	3.96	4.26	4.16	4.42
861	training	559	1,4-dimethylnaphthalene	571-58-4	Cc1ccc(C)c2ccccc12	3.754	4.37	3.79	3.71	3.96	4.26	4.16	4.37
862	training	560	1,5-dimethylnaphthalene	571-61-9	Cc1cccc2c(C)cccc12	3.760	4.37	3.79	3.71	3.96	4.26	4.16	4.38
863	training	561	1,7-dimethylnaphthalene	575-37-1	Cc2ccc1cccc(C)c1c2	3.792	4.36	3.79	3.71	3.96	4.26	4.16	4.44
864	training	562	2,3-dimethylnaphthalene	581-40-8	Cc2cc1ccccc1cc2C	3.771	4.37	3.79	3.71	3.96	4.26	4.16	4.40
865	training	563	2,6-dimethylnaphthalene	581-42-0	Cc2ccc1cc(C)ccc1c2	3.722	4.37	3.79	3.71	3.96	4.26	4.16	4.31
866	training	564	1-ethylnaphthalene	1127-76-0	CCc1cccc2ccccc12	3.771	4.47	3.84	3.68	3.96	4.21	4.19	4.39
867	training	565	2-ethylnaphthalene	939-27-5	CCc1cc2ccccc2cc1	3.760	4.47	3.84	3.68	3.96	4.21	4.19	4.38
868	training	566	hydrazobenzene	122-66-7	N(Nc(cccc1)c1)c(cccc2)c2	2.976	2.88	4.56	3.07	3.33	3.06	3.72	2.94
869	training	567	p-benzidine	92-87-5	Nc(ccc(c(ccc(N)c1)c1)c2)c2	5.360	1.59	2.22	1.86	2.51	1.92	1.91	1.34
870	training	568	chlorfenvinphos	470-90-6	CCOP(=O)(OCC)OC(=CCl)c1ccc(Cl)cc1Cl	2.470	4.05	3.87	3.56	4.45	4.15	4.87	3.10
871	training	569	4-phenylcyclohexanone	4894-75-1	c1ccccc1C2CCC(=O)CC2	2.710	2.62	3.24	2.46	2.63	2.76	2.30	2.14
872	training	570	diethyl phthalate	84-66-2	O=C(OCC)c(c(ccc1)C(=O)OCC)c1	2.721	2.60	2.80	2.24	2.58	2.65	2.74	2.47
873	training	571	carbofuran	1563-66-2	O=C(Oc(c(OC(C1)(C)C)c1cc2)c2)NC	1.790	2.08	2.62	2.23	1.46	2.30	2.09	2.32
874	training	572	hexylbenzene	1077-16-3	c(cccc1)(c1)CCCCCC	4.380	5.27	4.51	4.60	4.14	5.00	5.20	5.52
875	test	286	hexamethylbenzene	87-85-4	c(c(c(c(c1C)C)C)C)(c1C)C	3.928	4.71	3.87	4.75	4.14	5.28	4.64	4.61
876	training	573	isoproturon	34123-59-6	CC(C)c1ccc(NC(=O)N(C)C)cc1	2.350	2.63	2.69	2.35	2.39	2.84	2.46	2.87
877	training	574	diazinon	333-41-5	O(P(OCC)(Oc(nc(nc1C)C(C)C)c1)=S)CC	2.360	4.45	3.95	3.14	2.43	3.86	3.63	3.81
878	training	575	cyclododecanone	830-13-7	O=C(CCCCCCCCCC1)C1	3.607	4.45	4.97	3.91	2.93	4.07	4.33	4.19
879	test	287	dodecanoic acid	143-07-7	O=C(O)CCCCCCCCCCC	3.879	5.13	4.66	4.54	4.16	5.00	5.29	4.20

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
880	training	576	dodecane	112-40-3	C(CCCCCCCCCC)C	4.695	6.42	6.02	5.85	5.40	6.23	5.90	6.10
881	training	577	1-dodecanol	112-53-8	OCCCCCCCCCCCC	4.168	5.36	5.07	4.62	4.38	4.77	5.40	5.13
882	test	288	diethylene glycol dibutyl ether	112-73-2	O(CCOCCOCCCC)CCCC	2.421	2.48	3.04	2.45	1.75	2.46	2.34	1.92
883	test	289	hexachlorophene	70-30-4	Oc(c(c(c(c1)Cl)Cl)Cc(c(c(cc2Cl)Cl)Cl)c2O)c1Cl	3.515	6.77	6.79	7.26	4.80	6.92	6.27	7.54
884	training	578	9H-fluoren-9-one	486-25-9	O=C(c(c(c1cccc2)ccc3)c3)c12	3.325	3.45	3.69	2.92	3.32	3.55	3.49	3.58
885	training	579	acridine	260-94-6	n(c(c(ccc1)cc2cccc3)c1)c23	3.227	3.51	3.68	3.35	3.05	3.32	3.48	3.40
886	training	580	fluorene	86-73-7	c(c(c(c1ccc2)c2)ccc3)(c3)C1	3.662	4.26	4.12	3.49	4.15	4.02	3.91	4.18
887	training	581	mantuamycin	21609-90-5	COP(=S)(Oc1cc(Cl)c(Br)cc1Cl)c2ccccc2	5.070	6.37	6.34	5.81	4.54	6.34	7.15	6.31
888	training	582	benzophenone	119-61-9	O=C(c(cccc1)c1)c(cccc2)c2	2.640	3.03	3.25	3.23	3.59	3.15	3.58	3.43
889	test	290	phenyl benzoate	93-99-2	O=C(Oc(cccc1)c1)c(cccc2)c2	3.330	3.38	3.39	3.26	3.52	3.04	3.52	3.59
890	test	291	N-phenylbenzamide	93-98-1	O=C(Nc(cccc1)c1)c(cccc2)c2	2.802	2.43	3.07	2.62	3.11	2.70	2.99	2.62
891	training	583	diphenylmethane	101-81-5	c(cccc1)(c1)Cc(cccc2)c2	3.629	4.33	3.71	3.81	4.15	4.02	3.99	4.14
892	training	584	4-methylbiphenyl	644-08-6	c1ccccc1c2ccc(C)cc2	3.896	4.42	3.98	3.84	4.15	4.30	4.40	4.63
893	training	585	4-phenoxyphenyl urea	78508-44-8	NC(=O)Nc1ccc(Oc2ccccc2)cc1	2.560	2.49	2.27	2.30	2.14	2.76	1.88	2.80
894	training	586	benzyl phenyl ether	946-80-5	c1ccccc1OCc2ccccc2	3.439	3.63	3.24	3.40	3.40	3.78	3.65	3.79
895	training	587	diphenylmethanol	91-01-0	OC(c(cccc1)c1)c(cccc2)c2	2.829	2.76	2.96	2.84	3.13	2.71	2.79	2.67
896	test	292	4-biphenylmethanol	3597-91-9	OCc1ccc(cc1)c2ccccc2	2.690	3.18	3.04	2.74	3.13	2.84	3.15	3.38
897	training	588	imazapyr	81334-34-1	n1cccc(C(=O)O)c1C2=NC(C)(C(C)C)C(=O)N2	2.350	1.52	0.53	1.29	1.22	1.57	0.88	1.21
899	training	589	trifluralin	1582-09-8	CCCN(CCC)c1c(cc(cc1N(=O)(=O))C(F)(F)F)N(=O)(=O)	4.490	5.09	4.28	4.47	4.21	5.31	4.50	5.34
900	training	590	3-phenyl-1-cyclohexyl urea	886-59-9	O=C(NC1CCCCC1)Nc2ccccc2	2.070	3.25	2.96	2.81	2.27	3.44	2.74	2.91
901	test	293	fenamiphos	22224-92-6	CCOP(=O)(NC(C)C)Oc1ccc(SC)c(C)c1	2.520	3.05	3.06	3.25	3.75	3.29	3.31	3.23
902	training	591	1-tridecanol	112-70-9	OCCCCCCCCCCCCC	4.543	5.71	5.53	5.08	4.65	5.26	4.92	5.67
903	test	294	anthraquinone	84-65-1	O=C(c(c(C(=O)c1cccc2)ccc3)c3)c12	3.221	2.83	3.52	2.81	2.89	3.34	3.11	3.39
904	test	295	anthracene	120-12-7	c(c(ccc1)cc(c2ccc3)c3)(c1)c2	3.858	4.56	4.34	3.65	4.33	4.35	4.55	4.45
905	test	296	diphenylacetylene	501-65-5	C(#Cc(cccc1)c1)c(cccc2)c2	3.977	4.17	4.76	3.94	4.33	4.02	4.62	4.78
906	test	297	phenanthrene	85-01-8	c(c(c(c(c1)ccc2)c2)ccc3)(c1)c3	3.770	4.55	4.34	3.65	4.33	4.35	4.55	4.46
907	test	298	2-anthracenamine	613-13-8	Nc3ccc2cc1ccccc1cc2c3	4.480	3.69	3.62	2.90	3.58	3.43	3.74	3.48
908	training	592	trans-stilbene	103-30-0	c(cccc1)(c1)C=Cc(cccc2)c2	3.994	4.58	4.01	3.82	4.33	4.52	4.66	4.81
909	training	593	1-methylfluorene	1730-37-6	c1ccc2c3ccccc3Cc2c1C	4.081	4.56	4.43	3.98	4.41	4.56	4.34	4.97
910	training	594	2-phenylacetophenone	451-40-1	O=C(c(cccc1)c1)Cc(cccc2)c2	3.107	3.15	3.18	3.27	3.58	3.38	3.39	3.18
911	training	595	benzyl benzoate	120-51-4	O=C(OCc(cccc1)c1)c(cccc2)c2	3.537	3.43	3.32	3.27	3.51	3.54	3.67	3.97
912	training	596	1,2-diphenylethane	103-29-7	c(cccc1)(c1)CCc(cccc2)c2	3.934	4.74	3.93	4.26	4.41	4.74	4.37	4.79
913	training	597	4,4′-dimethylbiphenyl	613-33-2	c1cc(C)ccc1c2ccc(C)cc2	4.146	4.97	4.29	4.32	4.41	4.85	4.84	5.09
914	training	598	dibenzyl ether	103-50-4	O(Cc(cccc1)c1)Cc(cccc2)c2	3.178	3.42	3.18	3.22	3.39	3.48	3.44	3.31
915	test	299	triadimenol	55219-65-3	c1cc(Cl)ccc1OC(n2ncnc2)C(O)C(C)(C)C	1.950	2.88	2.86	3.25	3.41	2.95	3.06	3.08
916	test	300	alachlor	15972-60-8	CCc1cccc(CC)c1N(COC)C(=O)CCl	2.480	3.02	3.36	3.64	3.18	3.37	2.98	3.52
917	training	599	octylbenzene	2189-60-8	c(cccc1)(c1)CCCCCCCC	4.804	6.46	5.44	5.51	5.58	5.98	6.34	6.60
918	training	600	tetradecanoic acid	544-63-8	O=C(O)CCCCCCCCCCCCC	4.695	6.10	5.59	5.45	4.70	5.98	5.17	5.28
919	test	301	tetradecane	629-59-4	C(CCCCCCCCCCCC)C	5.294	7.70	6.95	6.76	5.93	7.22	6.83	7.18
920	training	601	1-tetradecanol	112-72-1	OCCCCCCCCCCCCCC	4.657	6.21	5.99	5.53	4.91	5.75	5.27	6.21

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
921	training	602	9-anthracenecarboxylic acid	723-62-6	O=C(O)c(c(c(ccc1)cc2cccc3)c1)c23	2.540	3.84	3.87	3.25	3.69	4.23	4.17	3.68
922	training	603	2-phenyl-1H-indene-1,3(2H)-dione	83-12-5	c1ccc2C(=O)C(c3ccccc3)C(=O)c2c1	2.955	3.10	3.29	2.82	3.15	2.81	2.65	2.90
923	training	604	2-methylanthracene	613-12-7	Cc1cc2cc3ccccc3cc2cc1	4.179	5.06	4.66	4.13	4.59	4.89	4.99	5.10
924	training	605	9-methylanthracene	779-02-2	c(c(c(c(c1ccc2)c2)C)ccc3)(c3)c1	4.135	5.06	4.66	4.13	4.59	4.89	4.99	5.07
925	training	606	1-methylphenanthrene	832-69-9	c1ccc2c3cccc(C)c3ccc2c1	4.173	5.05	4.66	4.13	4.59	4.89	4.99	5.12
926	test	302	9-anthracenemethanol	1468-95-7	OCc(c(c(ccc1)cc2cccc3)c1)c23	3.430	3.55	3.72	3.04	3.57	3.43	3.74	3.27
927	test	303	prochloraz	67747-09-5	Clc1cc(Cl)cc(Cl)c1OCCN(CCC)C(=O)n2cnc c2	4.130	3.78	4.28	4.41	3.37	4.13	4.29	4.60
928	training	607	bisphenol a	80-05-7	Oc(ccc(c1)C(c(ccc(O)c2)c2)(C)C)c1	3.183	3.81	3.46	3.73	3.31	3.64	4.10	3.32
929	training	608	imazamethabenz	100728-84-5	OC(=O)c1cc(C)ccc1C2=NC(C)(C(C)C)C(=O)N2	2.040	2.26	1.82	2.50	2.51	3.30	2.48	2.31
930	training	609	nicosulfuron	111991-09-4	CN(C)C(=O)c1cccnc1S(=O)(=O)NC(=O)Nc2 nc(OC)cc(OC)n2	2.260	0.59	0.55	0.69	0.39	-1.15	0.71	0.55
931	training	610	metalaxyl	57837-19-1	COCC(=O)N(C(C)C(=O)OC)c1c(C)cccc1C	1.660	1.47	2.08	2.38	1.91	1.70	1.58	1.65
932	training	611	metolachlor	51218-45-2	CCc1cccc(C)c1N(C(C)COC)C(=O)CCl	2.200	3.37	3.52	3.58	3.03	3.24	2.84	3.13
933	training	612	nonylbenzene	1081-77-2	c(cccc1)(c1)CCCCCCCCC	5.245	7.00	5.90	5.97	5.84	6.47	6.91	7.14
934	test	304	2,6-di-tert-butyl-p-cresol	128-37-0	Oc(c(cc(c1)C)C(C)(C)C)c1C(C)(C)C	4.151	5.25	5.03	4.85	4.18	5.03	5.26	5.27
935	training	613	fluoranthene	206-44-0	c(c(ccc1)ccc2)(c1c(c3ccc4)c4)c23	4.135	5.04	5.05	3.94	4.76	4.93	5.15	5.16
936	test	305	pyrene	129-00-0	c(c(c(cc1)ccc2)c2cc3)(c1ccc4)c34	4.800	5.19	5.03	3.94	4.76	4.93	5.15	4.50
937	training	614	9,10-dimethylphenanthrene	604-83-1	c(ccc1c(cccc2)c3c2)cc1c(c3C)C	4.472	5.40	4.97	4.62	4.84	5.44	5.43	5.15
938	training	615	dibutyl phthalate	84-74-2	O=C(OCCCC)c(c(ccc1)C(=O)OCCCC)c1	3.945	4.53	4.66	4.20	3.62	4.61	4.59	4.72
939	test	306	tebuconazole	107534-96-3	c1cc(Cl)ccc1CCC(O)(C(C)(C)C)Cn2ncnc2	2.670	3.60	3.48	3.63	4.01	3.89	3.38	3.70
940	test	307	decylbenzene	104-72-3	c(cccc1)(c1)CCCCCCCCCC	5.375	7.60	6.37	6.42	6.09	6.96	6.42	7.69
941	training	616	hexadecanoic acid	57-10-3	O=C(O)CCCCCCCCCCCCCCC	5.277	7.23	6.52	6.37	5.21	6.96	6.09	6.37
942	training	617	11H-benzo[a]fluorene	238-84-6	C3c1ccccc1c4ccc2ccccc2c34	4.315	5.46	5.30	4.40	5.00	5.19	5.17	5.68
943	training	618	11H-benzo[b]fluorene	243-17-4	c1ccc2c3cc4ccccc4cc3Cc2c1	4.505	5.31	5.30	4.40	5.00	5.19	5.17	5.77
944	training	619	ciprofloxacin	85721-33-1	C1CNCCN1c2cc3N(C4CC4)C=C(C(=O)O)C (=O)c3cc2F	4.790	-0.57	0.13	1.41	1.67	0.00	1.94	-1.08
945	test	308	morphine	57-27-2	C1=CC2C(N(C)C5)Cc3ccc(O)c4c3C2(C5)C(O4)C1O	1.829	0.99	1.53	1.39	1.93	0.72	0.76	0.76
946	training	620	napropamide	15299-99-7	CCN(CC)C(=O)C(C)Oc1cccc2ccccc12	2.540	3.43	3.62	3.28	3.10	3.33	3.89	3.36
947	test	309	undecylbenzene	6742-54-7	c1ccccc1CCCCCCCCCCC	5.805	8.02	6.83	6.88	6.33	7.45	6.78	8.23
948	test	310	chrysene	218-01-9	c1ccc2ccc3c4ccccc4ccc3c2c1	4.494	5.71	5.53	4.55	5.16	5.52	5.82	5.73
949	test	311	benz[a]anthracene	56-55-3	c(c(c(c(c1)ccc2)c2)cc(c3ccc4)c4)(c1)c3	4.592	5.72	5.53	4.55	5.16	5.52	5.82	5.79
950	test	312	naphthacene	92-24-0	c(c(cc(c1ccc2)c2)cc(c3ccc4)c4)(c1)c3	4.510	5.71	5.53	4.55	5.16	5.52	5.82	5.90
951	test	313	triphenylene	217-59-4	c1ccc2c3ccccc3c4ccccc4c2c1	4.364	5.77	5.53	4.55	5.16	5.52	5.82	5.49
952	training	621	2,2′-biquinoline	119-91-5	n(c(c(ccc1)cc2)c1)c2c(nc(c(ccc3)c4)c3)c4	4.260	4.31	4.41	4.58	3.73	4.06	4.35	4.31
953	test	314	6-chrysenamine	2642-98-0	Nc3cc2c1ccccc1ccc2c4ccccc34	5.580	4.81	4.80	3.81	4.42	4.60	5.00	4.98
954	training	622	p-terphenyl	92-94-4	c(c(cccc1)c1)(ccc(c(cccc2)c2)c3)c3	4.657	5.70	5.34	4.87	5.16	5.52	5.90	5.64
955	training	623	triphenylamine	603-34-9	c1ccccc1N(c2ccccc2)c3ccccc3	4.500	5.03	7.54	5.16	4.95	5.06	5.36	5.74
956	training	624	triphenyl phosphate	115-86-6	O=P(Oc(cccc1)c1)(Oc(cccc2)c2)Oc(cccc3)c3	3.874	4.16	4.34	4.94	4.79	4.70	5.06	4.59
957	training	625	triphenylphosphine	603-35-0	c(P(c(cccc1)c1)c(cccc2)c2)(cccc3)c3	4.472	5.46	5.67	5.71	5.97	5.02	5.50	4.61
958	training	626	dicumyl peroxide	80-43-3	O(OC(c(cccc1)c1)(C)C)C(c(cccc2)c2)(C)C	4.369	5.43	4.86	5.37	4.31	5.88	4.84	4.29

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
959	training	627	dodecylbenzene	123-01-3	c(cccc1)(c1)CCCCCCCCCCCC	6.083	8.38	7.29	7.33	6.57	7.94	7.35	8.77
960	training	628	linolenic acid	463-40-1	O=C(O)CCCCCCCC=CCC=CCC=CCC	4.891	6.62	6.38	5.95	4.48	7.30	5.85	5.86
961	test	315	linoleic acid	60-33-3	O=C(O)CCCCCCCC=CCC=CCCCCC	5.212	7.06	6.73	6.39	4.57	7.51	6.18	6.85
962	training	629	oleic acid 2027-47-6	112-80-1	O=C(O)CCCCCCCC=CCCCCCCCC	5.533	7.68	7.09	6.84	5.58	7.73	6.50	6.52
963	training	630	octadecanoic acid	57-11-4	O=C(O)CCCCCCCCCCCCCCCCCC	5.854	8.02	7.45	7.28	5.69	7.94	7.02	7.45
964	test	316	fluridone	59756-60-4	CN1C=C(c2ccccc2)C(=O)C(c3cc(C(F)(F)F)cc3)=C1	3.010	4.11	2.91	3.64	3.98	4.48	5.38	3.16
965	training	631	triphenylmethanol	76-84-6	OC(c(cccc1)c1)(c(cccc2)c2)c(cccc3)c3	3.379	4.31	3.63	4.28	4.38	4.38	4.75	3.68
966	training	632	pencycuron	66063-05-6	c1ccccc1NC(=O)N(C2CCCC2)Cc3ccc(Cl)cc3	3.330	4.77	4.85	4.80	4.00	5.51	4.65	4.82
967	training	633	enrofloxacin	93106-60-6	CCN1CCN(C4=C(C=C3C(C(C(O)=O)=CN(C3=C4)C2CC2)=O)F)CC1	4.850	0.58	0.85	2.30	2.13	0.70	2.61	-0.25
968	training	634	tridecylbenzene	123-02-4	c(cccc1)(c1)CCCCCCCCCCCCC	6.469	8.63	7.76	7.79	6.80	8.43	7.70	9.31
969	training	635	perylene	198-55-0	c(c(ccc1)ccc2)(c1c(c(c(cc3)ccc4)c45)c3)c25	4.777	6.34	6.23	4.85	5.55	6.11	6.41	5.75
970	training	636	benzo[a]pyrene	50-32-8	c(c(c(cc1)ccc2)c2cc3)(c3cc(c4ccc5)c5)c14	4.750	6.39	6.22	4.85	5.55	6.11	6.41	5.97

Conclusão

Mol ID	Status	pos	Nome	CAS	SMILES	Exp logKoc	ALOGPs	AC_logP	ALOGP	MLOGP	KOWWIN	XLOGP2	XLOGP3
971	test	317	13H-dibenzo[a,i]carbazole	239-64-5	c1ccc4c(c1)ccc5c3ccc2ccccc2c3nc45	6.100	6.10	5.80	5.14	4.53	5.58	6.06	6.40
973	training	637	7,12-dimethylbenz[a]anthracene	57-97-6	c(c(c(c(c1)ccc2)c2)c(c(c3ccc4)c4)C)(c3C)c1	5.290	6.61	6.16	5.53	5.63	6.62	6.69	5.80
974	training	638	5,8,11,14-eicosatetraenoic acid	506-32-1	CCCCCC=CCC=CCC=CCC=CCCCC(=O)O	5.174	6.80	6.95	6.41	4.85	8.07	6.45	6.29
975	training	639	eicosanoic acid	506-30-9	O=C(O)CCCCCCCCCCCCCCCCCCCC	6.431	8.53	8.38	8.19	6.15	8.93	7.95	8.53
976	training	640	1,2-dihydro-3-methylbenz[j]aceanthrylene	56-49-5	c(c(ccc1C)cc(c2ccc3cccc4)c34)(c1CC5)c25	6.100	6.49	6.34	5.64	5.85	7.05	6.58	6.42
978	training	641	cis-permethrin	61949-76-6	O=C(OCC2=CC=CC(OC3=CC=CC=C3)=C2)C1C(C)(C)C1C=C(Cl)Cl	3.190	6.24	5.91	5.44	4.58	7.43	6.13	6.50
982	test	318	trans-permethrin	61949-77-7	O=C(OCC2=CC=CC(OC3=CC=CC=C3)=C2)C1C(C)(C)C1C=C(Cl)Cl	3.190	6.24	5.91	5.44	4.58	7.43	6.13	6.50
985	test	319	strychnine	57-24-9	O=C(N(c(c(C1(C(N(C2)CC(C3C4C5OC6)=C6)C3)C2)ccc7)c7)C14)C5	4.140	1.68	1.48	1.15	2.90	1.85	0.56	1.93
986	test	320	benzo[ghi]perylene	191-24-2	c16cccc2ccc3ccc4ccc5cccc6c5c4c3c12	5.131	7.11	6.90	5.15	5.92	6.70	7.00	6.63
987	training	642	dibenz[a,h]anthracene	53-70-3	c(c(c(c(c1)ccc2)c2)cc(c3c(c(c4)ccc5)c5)c4)(c1)c3	6.070	6.93	6.71	5.46	5.92	6.70	7.09	6.50
988	test	321	coronene	191-07-1	c1cc3ccc4ccc5ccc6ccc7ccc1c2c7c6c5c4c23	4.668	7.26	7.59	5.45	6.28	7.28	7.60	7.24
989	training	643	dioctyl phthalate	117-84-0	O=C(OCCCCCCCC)c(c(ccc1)C(=O)OCCCCCCCC)c1	5.511	7.76	8.37	7.85	6.34	8.54	8.09	9.05

Nota: Adaptado de Shao et al. (2014).

Fórmulas estatísticas usadas no QSARINS

Tabela 2 Fórmulas usadas no ajuste (fitting)

Estatística	Definição	Equações e termos
$RMSE_{tr}$	Raiz do erro quadrado médio do conjunto de treinamento	$RMSE_{tr} = SDEC = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$ y_i = variável dependente observada $\hat{y}_i$ = variável dependente calculada
RSS_{tr}	Soma dos quadrados dos resíduos do conjunto de treinamento	$RSS_{tr} = \sum (y_i - \hat{y}_i)^2$
R^2	Coeficiente de determinação	$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS}$ $MSS = \sum_i (\hat{y}_i - \bar{y})^2$ $TSS = \sum_i (y_i - \bar{y})^2$ $RSS = \sum_i (y_i - \hat{y}_i)^2$ $\bar{y}$ = valor médio da variável dependente
R^2_{adj}	Coeficiente de determinação ajustado	$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} =$ $1 - (1 - R^2) \cdot \left(\frac{n-1}{n-p}\right)$ *n*= número de objetos *p* = número de variáveis preditoras
F	Estatística F	$F = \frac{MSS/(p)}{RSS/(n-p-1)}$
CCC_{tr}	Coeficiente de concordância da correlação do conjunto de treinamento	$CCC_{tr} = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$

Notas: Adaptado de Gramatica et al. (2013).

Tabela 3 Fórmulas usadas na validação cruzada (cross validation)

Estatística	Definição	Equações e termos
$PRESS_{cv}$	Soma dos quadrados dos resíduos da validação cruzada	$PRESS_{CV} = \sum (y_i - \hat{y}_{i/i})^2$ y_i = resposta observada para o *i-ésimo* objeto $\hat{y}_{i/i}$ = resposta do *i-ésimo* objeto estimado, usando um modelo obtido sem usar o *i-ésimo* objeto
$RMSE_{cv}$	Raiz do erro quadrado médio da validação cruzada	$RMSE_{CV} = SDEP = \sqrt{\frac{\sum (y_i - \hat{y}_{i/i})^2}{n}}$ *n* = número de objetos
Q^2_{LOO}	Coeficiente de determinação da validação LOO (*Leave-One-Out*)	$Q^2_{LOO} = 1 - \frac{PRESS_{CV}}{TSS}$ $TSS = \sum (y_i - \bar{y})^2$
CCC_{cv}	Coeficiente de concordância da correlação da validação cruzada	$CCC_{CV} = \frac{2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_{i/i} - \bar{\hat{y}})}{\sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_{i/i} - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$ $\bar{\hat{y}}$ = media de todos os $\hat{y}_{i/i}$

Tabela 4 Fórmulas usadas na validação externa (external validation)

Estatística	Definição	Equações e termos
$PRESS_{ext}$	Soma dos quadrados dos resíduos do conjunto de teste (validação externa)	$PRESS_{EXT} = \sum (y_i - \hat{y}_i)^2$ y_i = resposta externa observada para o *i-ésimo* objeto $\hat{y}_i$ = resposta externa predita usando o modelo
$RMSE_{ext}$	Raiz do erro quadrado médio do conjunto de teste	$RMSE_{EXT} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}}$ símbolos como acima
CCC_{ext}	Coeficiente de concordância da correlação do conjunto de teste	$CCC_{EXT} = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2}$ $\bar{\hat{y}}$ = media de todos os $\hat{y}_i$
R^2_{ext}	Coeficiente de determinação do conjunto de teste	$R^2_{EXT} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
$\overline{r_m^2}$	Critério de Roy: média	$\overline{r_m^2} = \frac{r_m^2 + r_m'^2}{2}$ $r_m^2 = r^2\left(1 - \sqrt{R^2_{ext} - R_0^2}\right)$ $r_m'^2 = r^2\left(1 - \sqrt{R^2_{ext} - R_0'^2}\right)$
Δr_m^2	Critério de Roy: delta	$\Delta r_m^2 = \left\lvert r_m^2 - r_m'^2 \right\rvert$ símbolos como acima

APÊNDICE B TABELAS DO ARTIGO 2

Tabela S1 Nomes, classes químicas, números CAS, valores experimentais de log K_{oc} e valores de log P obtidos pelo algoritmo ALOGPs para os compostos dos conjuntos de treinamento e de teste

Continua

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
1	training	1	bromotrifluoromethane	Alcano halogenado	75-63-8	2,389	1,55
2	training	2	carbon tetrabromide	Alcano halogenado	558-13-4	3,237	3,30
3	training	3	chlorotrifluoromethane	Alcano halogenado	75-72-9	2,275	1,80
4	training	4	dichlorodifluoromethane	Alcano halogenado	75-71-8	2,552	2,06
5	training	5	trichlorofluoromethane	Alcano halogenado	75-69-4	2,753	2,25
6	test	1	carbon tetrachloride	Alcano halogenado	56-23-5	2,27	2,64
7	training	6	carbon tetrafluoride	Alcano halogenado	75-73-0	2,019	1,75
8	training	7	bromoform	Alcano halogenado	75-25-2	2,672	2,50
9	test	2	chlorodifluoromethane	Alcano halogenado	75-45-6	1,965	0,98
10	training	8	dichlorofluoromethane	Alcano halogenado	75-43-4	2,22	1,28
11	test	3	chloroform	Alcano halogenado	67-66-3	1,65	1,67
12	test	4	fluoroform	Alcano halogenado	75-46-7	1,725	0,97
13	training	9	bromochloromethane	Alcano halogenado	74-97-5	2,144	1,27
14	training	10	dibromomethane	Alcano halogenado	74-95-3	2,628	1,48
15	training	11	chlorofluoromethane	Alcano halogenado	593-70-4	1,654	0,62
16	training	12	dichloromethane	Alcano halogenado	75-09-2	2,057	1,12
17	training	13	difluoromethane	Alcano halogenado	75-10-5	1,486	0,29
18	training	14	diiodomethane	Alcano halogenado	75-11-6	2,737	2,25
19	training	15	formaldehyde	Compostos carbonílicos	50-00-0	1,567	-0,68
20	training	16	formic acid	Ácido orgânico	64-18-6	1,083	-0,47
21	training	17	methyl bromide	Alcano halogenado	74-83-9	0,79	0,68
22	training	18	methyl chloride	Alcano halogenado	74-87-3	1,872	0,67
23	test	5	methyl fluoride	Alcano halogenado	593-53-3	1,654	0,41
24	test	6	methyl iodide	Alcano halogenado	74-88-4	1,04	1,20
25	training	19	formamide	Amida	75-12-7	0,556	-1,53
26	training	20	nitromethane	Nitroalcano	75-52-5	1,197	-0,17
27	training	21	methane	Alcano	74-82-8	1,97	-1,32
28	training	22	methyl alcohol	Álcool	67-56-1	0,974	-1,38
29	test	7	methylamine	Amina	74-89-5	1,067	-1,06
30	test	8	carbon disulfide	Organossulfurado	75-15-0	2,541	2,25
31	test	9	1,2-dichloro-1,1,2,2-tetrafluoroethane	Alcano halogenado	76-14-2	2,911	2,57
32	training	23	1,1,2-trichloro-1,2,2-trifluoroethane	Alcano halogenado	76-13-1	3,096	3,03
33	training	24	tetrachloroethylene	Alceno halogenado	127-18-4	2,31	3,13
34	training	25	hexachloroethane	Alcano halogenado	67-72-1	3,553	3,93
35	training	26	hexafluoroethane	Alcano halogenado	76-16-4	2,465	2,46
36	test	10	halothane	Alcano halogenado	151-67-7	2,628	2,50
37	training	27	trichloroethylene	Alceno halogenado	79-01-6	2,15	2,45
38	training	28	trichloroacetaldehyde	Compostos carbonílicos	75-87-6	1,916	1,38
39	test	11	pentachloroethane	Alcano halogenado	76-01-7	2,949	3,21
40	test	12	acetylene	Alcenos e alcinos	74-86-2	1,578	-0,03
41	training	29	1,1-dichloroethylene	Alceno halogenado	75-35-4	2,536	1,97
42	test	13	cis-1,2-dichloroethylene	Alceno halogenado	156-59-2	2,389	1,85
43	training	30	trans-1,2-dichloroethylene	Alceno halogenado	156-60-5	2,427	1,85
44	training	31	dichloroacetic acid	Ácido orgânico	79-43-6	1,877	0,99
45	training	32	2,2,2-trichloroacetamide	Amida	594-65-0	1,943	0,98
46	training	33	1,1,2,2-tetrachloroethane	Alcano halogenado	79-34-5	2,677	2,57
47	training	34	1,1-difluoroethylene	Alceno halogenado	75-38-7	2,052	1,56
48	test	14	trifluoroacetamide	Amida	354-38-1	1,442	0,08
49	training	35	vinyl bromide	Alceno halogenado	593-60-2	2,231	1,19

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
50	training	36	bromoacetic acid	Ácido orgânico	79-08-3	1,6	0,53
51	training	37	vinyl chloride	Alceno halogenado	75-01-4	2,128	1,43
52	training	38	chloroacetic acid	Ácido orgânico	79-11-8	1,497	0,18
53	training	39	1,1,1-trichloroethane	Alcano halogenado	71-55-6	2,01	2,45
54	training	40	1,1,2-trichloroethane	Alcano halogenado	79-00-5	1,8	2,02
55	test	15	2,2,2-trichloroethanol	Álcool	115-20-8	2,111	1,23
56	test	16	2,2,2-trifluoroethanol	Álcool	75-89-8	1,6	0,61
57	test	17	acetonitrile	Nitrila	75-05-8	1,192	-0,04
58	test	18	ethylene	Alcenos e alcinos	74-85-1	1,992	0,90
59	training	41	1,2-dibromoethane	Alcano halogenado	106-93-4	2,443	2,08
60	test	19	1,1-dichloroethane	Alcano halogenado	75-34-3	1,49	1,72
61	training	42	1,2-dichloroethane	Alcano halogenado	107-06-2	1,65	1,48
63	test	20	dichloroethane	Alcano halogenado	1300-21-6	1,785	1,48
64	training	43	1,2-diiodoethane	Alcano halogenado	624-73-7	2,851	2,72
65	training	44	acetaldehyde	Compostos carbonílicos	75-07-0	1,622	-0,01
66	training	45	ethylene oxide	Éter	75-21-8	1,214	-0,47
67	test	21	acetic acid	Ácido orgânico	64-19-7	1,285	-0,12
68	test	22	methyl formate	Éster	107-31-3	1,393	-0,31
69	training	46	bromoethane	Alcano halogenado	74-96-4	2,247	1,64
70	test	23	ethyl chloride	Alcano halogenado	75-00-3	2,155	1,47
71	training	47	2-chloroethanol	Álcool	107-07-3	1,393	0,00
72	training	48	ethyl iodide	Alcano halogenado	75-03-6	2,465	2,29
73	training	49	acetamide	Amida	60-35-5	0,692	-1,10
74	training	50	N-methylformamide	Amida	123-39-7	0,849	-1,31
75	test	24	nitroethane	Nitroalcano	79-24-3	1,475	0,45
76	training	51	ethane	Alcano	74-84-0	2,362	1,44
77	test	25	ethyl alcohol	Álcool	64-17-5	1,214	-0,40
78	training	52	dimethyl ether	Éter	115-10-6	1,431	-0,16
79	training	53	dimethyl sulfoxide	Organossulfurado	67-68-5	0,643	-1,09
80	training	54	ethylene glycol	Álcool	107-21-1	0,637	-1,53
81	training	55	dimethyl sulfone	Organossulfurado	67-71-0	0,61	-0,95
82	training	56	dimethyl sulfate	Organossulfurado	77-78-1	2,008	-0,60
83	training	57	dimethyl disulfide	Organossulfurado	624-92-0	2,34	1,15
84	test	26	ethylamine	Amina	75-04-7	1,306	-0,20
85	test	27	dimethylamine	Amina	124-40-3	2,72	-0,53
86	test	28	monoethanolamine	Amina	141-43-5	0,664	-1,53
87	training	58	ethylenediamine	Amina	107-15-3	0,267	-1,77
88	training	59	cyanogen	Nitrila	460-19-5	1,415	-0,65
89	training	60	hexafluoroacetone	Compostos carbonílicos	684-16-2	2,171	1,77
90	training	61	malononitrile	Nitrila	109-77-3	1,051	-0,84
91	test	29	acrylonitrile	Nitrila	107-13-1	1,513	0,20
92	training	62	oxazole	Heterociclo aromático	288-42-6	1,442	-0,09
93	training	63	thiazole	Heterociclo aromático	288-47-1	1,616	0,89
94	test	30	methylacetylene	Alcenos e alcinos	74-99-7	1,888	0,92
95	training	64	allene	Alcenos e alcinos	463-49-0	2,166	1,67
96	training	65	cis-1,2-dichloropropene	Alceno halogenado	6923-20-2	2,481	2,10
97	training	66	imidazole	Heterociclo aromático	288-32-4	1,333	-0,21
98	training	67	1H-pyrazole	Heterociclo aromático	288-13-1	1,448	0,03
99	training	68	acrolein	Compostos carbonílicos	107-02-8	1,372	0,18
100	test	31	propargyl alcohol	Álcool	107-19-7	1,17	-0,70
101	test	32	acrylic acid	Ácido orgânico	79-10-7	1,567	0,46
102	test	33	3-bromo-1-propene	Alceno halogenado	106-95-6	2,351	1,98
103	test	34	2-chloro-1-propene	Alceno halogenado	557-98-2	2,465	1,88
104	training	69	α-epichlorohydrin	Éter	106-89-8	1,54	0,35
105	training	70	1,2,3-trichloropropane	Alcano halogenado	96-18-4	2,612	2,29
106	test	35	propionitrile	Nitrila	107-12-0	1,464	-0,01
107	training	71	acrylamide	Amida	79-06-1	0,953	-0,65
108	test	36	lactonitrile	Nitrila	78-97-7	0,866	-0,65
109	training	72	nitroglycerine	Outros compostos	55-63-0	2,258	1,25
110	test	37	cyclopropane	Alcano	75-19-4	2,313	1,56
111	test	38	propylene	Alcenos e alcinos	115-07-1	2,34	1,68
112	training	73	1,2-dichloropropane	Alcano halogenado	78-87-5	2,465	2,13

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
113	training	74	allyl alcohol	Álcool	107-18-6	1,469	-0,03
114	training	75	propionaldehyde	Compostos carbonílicos	123-38-6	1,698	0,31
115	test	39	acetone	Compostos carbonílicos	67-64-1	1,246	-0,29
116	training	76	1,2-propylene oxide	Éter	75-56-9	1,393	0,04
117	training	77	1,3-propylene oxide	Éter	503-30-0	1,301	0,05
119	training	78	propanoic acid	Ácido orgânico	79-09-4	1,557	0,31
120	training	79	ethyl formate	Ácido orgânico	109-94-4	1,502	0,38
121	training	80	methyl acetate	Ácido orgânico	79-20-9	1,475	0,18
122	test	40	3-mercaptopropionic acid	Ácido orgânico	107-96-0	1,611	0,34
123	test	41	lactic acid	Ácido orgânico	50-21-5	0,985	-0,79
124	test	42	trioxane	Heterociclo	110-88-3	1,143	-0,95
125	training	81	1-bromopropane	Alcano halogenado	106-94-5	2,519	2,18
126	test	43	2-bromopropane	Alcano halogenado	75-26-3	2,411	1,83
127	test	44	1-chloropropane	Alcano halogenado	540-54-5	2,487	2,09
128	training	82	2-chloropropane	Alcano halogenado	75-29-6	2,411	1,49
129	training	83	1-iodopropane	Alcano halogenado	107-08-4	2,737	2,65
130	training	84	2-iodopropane	Alcano halogenado	75-30-9	2,949	2,59
131	test	45	allylamine	Amina	107-11-9	1,393	-0,43
132	training	85	N,N-dimethylformamide	Amida	68-12-2	0,828	-0,77
133	training	86	N-methylacetamide	Amida	79-16-3	0,806	-1,06
134	training	87	1-nitropropane	Nitroalcano	108-03-2	1,85	0,91
135	training	88	2-nitropropane	Nitroalcano	79-46-9	1,883	0,71
136	training	89	propane	Alcano	74-98-6	2,661	2,19
137	training	90	glyphosate	Organofosforados	1071-83-6	3,46	-2,43
138	training	91	propyl alcohol	Álcool	71-23-8	1,513	0,21
139	training	92	isopropyl alcohol	Álcool	67-63-0	1,404	0,04
140	training	93	2-methoxyethanol	Álcool	109-86-4	0,958	-0,78
141	training	94	1,2-propanediol	Álcool	57-55-6	0,877	-1,10
142	test	46	1,3-propanediol	Álcool	504-63-2	0,811	-1,18
143	test	47	glycerol	Álcool	56-81-5	0,42	-1,93
144	test	48	propyl mercaptan	Organossulfurado	107-03-9	2,362	1,72
145	training	95	methyl ethyl sulfide	Organossulfurado	624-89-5	2,215	1,16
146	training	96	propylamine	Amina	107-10-8	1,638	0,31
147	training	97	isopropylamine	Amina	75-31-0	1,518	-0,05
148	training	98	methylethylamine	Amina	624-78-2	1,459	0,13
149	training	99	trimethylamine	Amina	75-50-3	1,464	-0,14
150	training	100	1-amino-2-propanol	Amina	78-96-6	0,855	-1,03
151	test	49	3-amino-1-propanol	Amina	156-87-6	0,768	-1,01
152	training	101	methylethanolamine	Amina	109-83-1	0,866	-1,05
153	training	102	trimethyl phosphate	Organofosforados	512-56-1	1,023	-0,61
154	training	103	hexachloro-1,3-butadiene	Alcenos e alcinos	87-68-3	3,977	4,86
155	test	50	succinonitrile	Nitrila	110-61-2	0,838	-0,75
156	training	104	pyrimidine	Heterociclo aromático	289-95-2	1,159	-0,21
157	test	51	furan	Heterociclo aromático	110-00-9	2,106	1,24
158	test	52	fumaric acid	Ácido orgânico	110-17-8	1,627	0,21
159	training	105	maleic acid	Ácido orgânico	110-16-7	1,116	0,21
160	training	106	thiophene	Heterociclo aromático	110-02-1	2,362	1,89
161	training	107	methacrylonitrile	Nitrila	126-98-7	1,747	0,91
162	test	53	vinylacetonitrile	Nitrila	109-75-1	1,595	0,61
163	training	108	pyrrole	Heterociclo aromático	109-97-7	1,785	0,76
164	training	109	methyl cyanoacetate	Ácido orgânico	105-34-0	1,121	-0,10
165	training	110	dimethylacetylene	Alcenos e alcinos	503-17-3	2,171	1,70
166	training	111	1,3 butadiene	Alcenos e alcinos	106-99-0	2,46	1,94
167	training	112	2,5-dihydrofuran	Heterociclo	1708-29-8	1,627	0,48
168	training	113	γ-butyrolactone	Éster	96-48-0	1,029	-0,11
169	training	114	methacrylic acid	Ácido orgânico	79-41-4	1,883	0,63
170	test	54	methyl acrylate	Ácido orgânico	96-33-3	1,812	0,67
171	training	115	vinyl acetate	Ácido orgânico	108-05-4	1,774	0,83
172	training	116	succinic acid	Ácido orgânico	110-15-6	1,056	-0,53
173	training	117	butyronitrile	Nitrila	109-74-0	1,703	0,59
174	training	118	isobutyronitrile	Nitrila	78-82-0	1,627	0,50
175	training	119	2-pyrrolidone	Amida	616-45-5	0,915	-0,90

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
176	training	120	1-butene	Alcenos e alcinos	106-98-9	2,683	2,21
177	training	121	cis-2-butene	Alcenos e alcinos	590-18-1	2,645	2,32
178	training	122	trans-2-butene	Alcenos e alcinos	624-64-6	2,634	2,32
179	training	123	isobutene	Alcenos e alcinos	115-11-7	2,655	1,87
180	training	124	bis(2-chloroethyl) ether	Éter	111-44-4	1,986	1,23
181	training	125	ethyl vinyl ether	Alcenos e alcinos	109-92-2	1,943	1,19
182	training	126	butyraldehyde	Compostos carbonílicos	123-72-8	1,856	1,10
183	training	127	methyl ethyl ketone	Compostos carbonílicos	78-93-3	1,535	0,41
184	test	55	tetrahydrofuran	Éter	109-99-9	1,627	0,35
185	training	128	butyric acid	Ácido orgânico	107-92-6	1,807	0,78
186	training	129	isobutyric acid	Ácido orgânico	79-31-2	1,888	0,78
187	training	130	propyl formate	Éster	110-74-7	1,829	0,93
188	test	56	ethyl acetate	Éster	141-78-6	1,774	0,74
189	training	131	methyl propanoate	Éster	554-12-1	1,834	0,68
190	training	132	1,4-dioxane	Heterociclo	123-91-1	1,149	-0,23
191	test	57	sulfolane	Organossulfurado	126-33-0	0,958	-0,65
192	test	58	1-bromobutane	Alcano halogenado	109-65-9	2,873	2,73
193	training	133	1-chlorobutane	Alcano halogenado	109-69-3	2,813	2,37
194	test	59	2-chlorobutane	Alcano halogenado	78-86-4	2,645	2,34
195	training	134	1-fluorobutane	Alcano halogenado	2366-52-1	2,781	1,79
196	test	60	1-iodobutane	Alcano halogenado	542-69-8	3,009	3,11
197	training	135	pyrrolidine	Heterociclo	123-75-1	1,627	0,16
198	training	136	N,N-dimethylacetamide	Amida	127-19-5	0,958	-0,59
199	training	137	morpholine	Heterociclo	110-91-8	0,398	-0,75
200	training	138	butanamide	Amida	541-35-5	1,263	-0,13
201	training	139	1-nitrobutane	Nitroalcano	627-05-4	2,177	1,49
202	training	140	butane	Alcano	106-97-8	2,949	2,81
203	training	141	piperazine	Heterociclo	110-85-0	0,741	-1,16
204	training	142	butanol	Álcool	71-36-3	1,834	0,84
205	test	61	isobutanol	Álcool	78-83-1	1,79	0,60
206	training	143	sec-butanol	Álcool	78-92-2	1,731	0,66
207	training	144	tert-butanol	Álcool	75-65-0	1,567	0,70
208	test	62	diethyl ether	Éter	60-29-7	1,861	1,12
209	training	145	methyl propyl ether	Éter	557-17-5	2,035	0,90
210	test	63	1,2-dimethoxyethane	Éter	110-71-4	1,263	0,03
211	test	64	2-ethoxyethanol	Álcool	110-80-5	1,203	-0,28
212	training	146	1,4-butanediol	Álcool	110-63-4	0,925	-0,63
213	training	147	diethyl sulfate	Organossulfurado	64-67-5	1,997	-0,29
214	test	65	butyl mercaptan	Organossulfurado	109-79-5	2,617	2,51
215	test	66	diethyl sulfide	Organossulfurado	352-93-2	2,438	2,46
216	training	148	butylamine	Amina	109-73-9	1,845	0,85
217	training	149	isobutylamine	Amina	78-81-9	1,774	0,54
218	training	150	tert-butylamine	Amina	75-64-9	1,595	0,81
219	test	67	diethylamine	Amina	109-89-7	1,693	0,76
220	test	68	diethanolamine	Amina	111-42-2	0,599	-1,41
221	training	151	hexachlorocyclopentadiene	Alcenos e alcinos	77-47-4	4,119	4,85
222	training	152	furfural	Heterociclo aromático	98-01-1	1,6	0,43
223	test	69	pyridine	Heterociclo aromático	110-86-1	1,731	0,70
224	training	153	glutaronitrile	Nitrila	544-13-8	0,985	-0,49
225	training	154	2-methylfuran	Heterociclo aromático	534-22-5	2,383	1,75
226	test	70	furfuryl alcohol	Heterociclo aromático	98-00-0	1,529	0,25
227	training	155	2-methylthiophene	Heterociclo aromático	554-14-3	2,645	2,30
228	test	71	3-methylthiophene	Heterociclo aromático	616-44-4	2,65	2,28
229	test	72	N-methylpyrrole	Heterociclo aromático	96-54-8	2,035	1,31
230	training	156	isoprene	Alcenos e alcinos	78-79-5	2,693	2,22
231	training	157	cis-1,3-pentadiene	Alcenos e alcinos	1574-41-0	2,683	2,65
232	training	158	trans-1,3-pentadiene	Alcenos e alcinos	2004-70-8	2,704	2,65
233	training	159	1,4-pentadiene	Alcenos e alcinos	591-93-5	2,726	2,39
234	training	160	1-pentyne	Alcenos e alcinos	627-19-0	2,454	2,13
235	training	161	acetylacetone	Compostos carbonílicos	123-54-6	1,595	-0,20
236	test	73	allyl acetate	Éster	591-87-7	1,905	1,03
237	training	162	ethyl acrylate	Éster	140-88-5	2,095	1,24

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
238	training	163	methyl methacrylate	Éster	80-62-6	2,128	1,10
239	training	164	2-hydroxyethyl acrylate	Éster	818-61-1	1,263	0,04
240	test	74	levulinic acid	Ácido orgânico	123-76-2	1,11	-0,14
241	training	165	glutaric acid	Ácido orgânico	110-94-1	1,219	-0,25
242	training	166	valeronitrile	Nitrila	110-59-8	1,888	1,10
243	training	167	N-methyl-2-pyrrolidone	Amida	872-50-4	1,17	-0,72
244	test	75	L-glutamic acid	Ácido orgânico	56-86-0	-0,63	-3,54
245	training	168	cyclopentane	Alcano	287-92-3	3,009	2,88
246	test	76	methyl propyl ketone	Compostos carbonílicos	107-87-9	1,834	0,87
247	training	169	diethyl ketone	Compostos carbonílicos	96-22-0	1,823	1,19
248	training	170	methyl isopropyl ketone	Compostos carbonílicos	563-80-4	1,682	0,78
249	test	77	2-methyltetrahydrofuran	Éter	96-47-9	2,383	0,96
250	test	78	tetrahydropyran	Éter	142-68-7	1,823	1,16
251	training	171	pentanoic acid	Ácido orgânico	109-52-4	2,133	1,34
252	training	172	3-methylbutanoic acid	Ácido orgânico	503-74-2	2,008	1,26
253	test	79	propyl acetate	Éster	109-60-4	2,052	1,28
254	training	173	ethyl propanoate	Éster	105-37-3	2,035	1,32
255	training	174	methyl butanoate	Éster	623-42-7	2,079	1,22
256	training	175	diethyl carbonate	Éster	105-58-8	2,035	0,86
257	training	176	1-bromopentane	Alcano halogenado	110-53-2	3,21	3,27
258	test	80	1-chloropentane	Alcano halogenado	543-59-9	2,862	3,12
259	training	177	2-chloro-2-methylbutane	Alcano halogenado	594-36-5	2,748	2,95
260	training	178	1-fluoropentane	Alcano halogenado	592-50-7	2,645	2,93
261	test	81	N-methylpyrrolidine	Heterociclo	120-94-5	1,877	0,54
262	training	179	piperidine	Heterociclo	110-89-4	1,834	0,97
263	training	180	1-nitropentane	Nitroalcano	628-05-7	2,47	2,00
264	training	181	pentane	Alcano	109-66-0	3,254	3,41
265	training	182	isopentane	Alcano	78-78-4	2,628	3,12
266	training	183	neopentane	Alcano	463-82-1	3,069	2,95
267	training	184	dimethoate	Organofosforados	60-51-5	2,56	1,21
268	training	185	1-pentanol	Álcool	71-41-0	2,198	1,47
269	training	186	2-pentanol	Álcool	6032-29-7	2,057	1,18
270	test	82	3-pentanol	Álcool	584-02-1	2,035	1,22
271	training	187	2-methyl-1-butanol	Álcool	137-32-6	2,079	1,24
272	training	188	3-methyl-1-butanol	Álcool	123-51-3	2,073	1,33
273	training	189	tert-pentyl-alcohol	Álcool	75-85-4	1,861	1,19
274	test	83	3-methyl-2-butanol	Álcool	598-75-4	2,073	0,89
275	training	190	2,2-dimethyl-1-propanol	Álcool	75-84-3	2,09	1,15
276	test	84	methyl tert-butyl ether	Éter	1634-04-4	1,888	1,53
277	training	191	pentaerythritol	Álcool	115-77-5	0,458	-1,92
278	training	192	pentylamine	Amina	110-58-7	2,188	1,39
279	training	193	hexachlorobenzene	Benzeno halogenado	118-74-1	4,49	5,70
280	training	194	hexafluorobenzene	Benzeno halogenado	392-56-3	2,764	2,33
281	test	85	pentachlorobenzene	Benzeno halogenado	608-93-5	4,113	5,22
282	test	86	pentachlorophenol	Fenóis	87-86-5	2,47	4,99
283	training	195	1,2,3,4-tetrachlorobenzene	Benzeno halogenado	634-66-2	3,52	4,62
284	training	196	1,2,3,5-tetrachlorobenzene	Benzeno halogenado	634-90-2	3,52	4,63
285	test	87	1,2,4,5-tetrachlorobenzene	Benzeno halogenado	95-94-3	3,72	4,61
286	training	197	2,3,4,5-tetrachlorophenol	Fenóis	4901-51-3	2,88	4,41
287	test	88	2,3,4,6-tetrachlorophenol	Fenóis	58-90-2	2,88	4,37
288	training	198	2,3,5,6-tetrachlorophenol	Fenóis	935-95-5	2,88	4,45
289	test	89	1-chloro-2,4-dinitrobenzene	Nitrobenzeno	97-00-7	2,557	2,29
290	test	90	1,2-dichloro-4-nitrobenzene	Nitrobenzeno	99-54-7	2,53	3,11
291	test	91	1,2,4-trichlorobenzene	Benzeno halogenado	120-82-1	3,11	4,08
292	test	92	1,2,3-trichlorobenzene	Benzeno halogenado	87-61-6	3,23	4,07
293	training	199	1,3,5-trichlorobenzene	Benzeno halogenado	108-70-3	2,85	4,08
295	training	200	2,3,4-trichlorophenol	Fenóis	15950-66-0	1,96	3,78
296	training	201	2,3,5-trichlorophenol	Fenóis	933-78-8	1,96	3,77
297	test	93	2,3,6-trichlorophenol	Fenóis	933-75-5	1,96	3,77
298	training	202	2,4,5-trichlorophenol	Fenóis	95-95-4	1,96	3,79
299	training	203	2,4,6-trichlorophenol	Fenóis	88-06-2	1,96	3,78
300	test	94	3,4,5-trichlorophenol	Fenóis	609-19-8	1,96	3,77

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
301	test	95	nitrapyrin	Heterociclo aromático	1929-82-4	2,24	3,87
302	training	204	1,3,5-trinitrobenzene	Nitrobenzeno	99-35-4	2,019	1,54
303	test	96	1-bromo-2-chlorobenzene	Benzeno halogenado	694-80-4	2,6	3,61
304	training	205	1-bromo-3-chlorobenzene	Benzeno halogenado	108-37-2	2,6	3,59
305	training	206	1-bromo-4-chlorobenzene	Benzeno halogenado	106-39-8	2,6	3,63
307	test	97	3-bromo-5-chlorophenol	Fenóis	56962-04-0	2,6	3,14
308	training	207	4-bromo-2-chlorophenol	Fenóis	3964-56-5	2,6	3,15
309	training	208	2-bromo-4-chlorophenol	Fenóis	695-96-5	2,6	3,12
310	training	209	1-bromo-2-nitrobenzene	Nitrobenzeno	577-19-5	2,42	2,59
311	test	98	1-bromo-3-nitrobenzene	Nitrobenzeno	585-79-5	2,42	2,61
312	training	210	1-bromo-4-nitrobenzene	Nitrobenzeno	586-78-7	2,42	2,66
313	training	211	m-dibromobenzene	Benzeno halogenado	108-36-1	3,417	3,73
314	training	212	m-chloronitrobenzene	Nitrobenzeno	121-73-3	2,715	2,49
315	training	213	o-chloronitrobenzene	Nitrobenzeno	88-73-3	2,596	2,48
316	test	99	p-chloronitrobenzene	Nitrobenzeno	100-00-5	2,677	2,56
317	test	100	o-dichlorobenzene	Benzeno halogenado	95-50-1	2,78	3,45
318	training	214	m-dichlorobenzene	Benzeno halogenado	541-73-1	2,78	3,45
319	training	215	p-dichlorobenzene	Benzeno halogenado	106-46-7	2,78	3,46
320	test	101	2,3-dichlorophenol	Fenóis	576-24-9	2,55	3,15
321	training	216	2,4-dichlorophenol	Fenóis	120-83-2	2,55	3,14
322	training	217	2,5-dichlorophenol	Fenóis	583-78-8	2,55	3,13
323	test	102	2,6-dichlorophenol	Fenóis	87-65-0	2,55	3,15
324	training	218	3,4-dichlorophenol	Fenóis	95-77-2	2,55	3,12
325	training	219	3,5-dichlorophenol	Fenóis	591-35-5	2,55	3,09
326	test	103	m-difluorobenzene	Benzeno halogenado	372-18-9	2,579	2,25
327	test	104	o-difluorobenzene	Benzeno halogenado	367-11-3	2,666	2,24
328	training	220	p-difluorobenzene	Benzeno halogenado	540-36-3	2,536	2,26
329	test	105	m-dinitrobenzene	Nitrobenzeno	99-65-0	2,188	1,70
330	training	221	o-dinitrobenzene	Nitrobenzeno	528-29-0	2,296	1,64
331	test	106	p-dinitrobenzene	Nitrobenzeno	100-25-4	2,171	1,70
332	training	222	bromobenzene	Benzeno halogenado	108-86-1	3,004	2,65
333	test	107	o-bromophenol	Fenóis	95-56-7	2,41	2,52
334	training	223	m-bromophenol	Fenóis	591-20-8	2,41	2,46
335	training	224	p-bromophenol	Fenóis	106-41-2	2,41	2,50
337	training	225	chlorobenzene	Benzeno halogenado	108-90-7	2,22	2,78
338	training	226	m-chlorophenol	Fenóis	108-43-0	1,82	2,35
339	training	227	o-chlorophenol	Fenóis	95-57-8	1,71	2,40
340	test	108	p-chlorophenol	Fenóis	106-48-9	1,85	2,37
341	training	228	3,4-dichloroaniline	Anilinas	95-76-1	0,67	2,74
342	training	229	2,3-dichloroaniline	Anilinas	608-27-5	0,67	2,73
343	test	109	2,4-dichloroaniline	Anilinas	554-00-7	0,67	2,73
344	training	230	2,5-dichloroaniline	Anilinas	95-82-9	0,67	2,72
345	training	231	2,6-dichloroaniline	Anilinas	608-31-1	0,67	2,74
346	test	110	3,5-dichloroaniline	Anilinas	626-43-7	0,67	2,71
347	training	232	fluorobenzene	Benzeno halogenado	462-06-6	2,612	2,18
348	training	233	iodobenzene	Benzeno halogenado	591-50-4	3,161	3,00
349	training	234	nitrobenzene	Nitrobenzeno	98-95-3	2,01	1,89
350	training	235	o-nitrophenol	Fenóis	88-75-5	2,06	1,91
351	test	111	m-nitrophenol	Fenóis	554-84-7	1,72	1,92
352	training	236	p-nitrophenol	Fenóis	100-02-7	2,72	1,93
353	test	112	benzene	Benzeno e Alquil benzeno	71-43-2	1,87	2,03
354	training	237	o-bromoaniline	Anilinas	615-36-1	1,96	2,14
355	training	238	m-bromoaniline	Anilinas	591-19-5	1,96	2,16
356	test	113	p-bromoaniline	Anilinas	106-40-1	1,96	2,10
357	training	239	m-chloroaniline	Anilinas	108-42-9	3,13	1,93
358	training	240	o-chloroaniline	Anilinas	95-51-2	3,13	1,93
359	test	114	p-chloroaniline	Anilinas	106-47-8	3,13	1,95
360	test	115	1α,2α,3ß,4α,5α,6ß-hexachlorocyclohexane	Alcano halogenado	58-89-9	3,41	3,94
361	test	116	1α,2ß,3α,4ß,5α,6ß-hexachlorocyclohexane	Alcano halogenado	319-85-7	3,12	3,94
362	training	241	1α,2α,3α,4ß,5α,6ß-hexachlorocyclohexane	Alcano halogenado	319-86-8	3,3	3,94
364	test	117	a-hexachlorocyclohexane	Alcano halogenado	319-84-6	3,3	3,94
366	test	118	m-nitroaniline	Anilinas	99-09-2	2,122	1,53

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
367	training	242	o-nitroaniline	Anilinas	88-74-4	2,383	1,43
368	test	119	p-nitroaniline	Anilinas	100-01-6	2,133	1,50
369	training	243	phenol	Fenóis	108-95-2	1,74	1,39
370	test	120	pyrocatechol	Fenóis	120-80-9	2,07	0,74
371	test	121	resorcinol	Fenóis	108-46-3	1,02	0,70
372	test	122	p-hydroquinone	Fenóis	123-31-9	1,698	0,71
373	test	123	phenyl mercaptan	Organossulfurado	108-98-5	2,748	2,26
374	training	244	aniline	Anilinas	62-53-3	1,867	0,89
375	test	124	2-methylpyridine	Heterociclo aromático	109-06-8	1,981	1,25
376	training	245	3-methylpyridine	Heterociclo aromático	108-99-6	2,03	1,11
377	test	125	4-methylpyridine	Heterociclo aromático	108-89-4	2,041	1,14
378	training	246	1,4-cyclohexadiene	Alcenos e alcinos	628-41-1	2,721	2,31
379	training	247	adiponitrile	Nitrila	111-69-3	1,203	-0,14
380	training	248	m-phenylenediamine	Anilinas	108-45-2	1,197	0,01
381	training	249	o-phenylenediamine	Anilinas	95-54-5	1,459	-0,08
382	training	250	p-phenylenediamine	Anilinas	106-50-3	1,214	-0,01
383	test	126	phenylhydrazine	Outros compostos	100-63-0	2,057	0,95
384	test	127	2-ethylfuran	Heterociclo aromático	3208-16-0	2,683	2,50
385	test	128	2-cyclohexen-1-one	Compostos carbonílicos	930-68-7	1,709	0,97
386	training	251	5-hexyn-2-one	Compostos carbonílicos	2550-28-9	1,693	1,13
387	test	129	ascorbic acid	Éster	50-81-7	0,485	-1,58
388	training	252	citric acid	Ácido orgânico	77-92-9	0,441	-1,33
389	training	253	cyclohexene	Alcenos e alcinos	110-83-8	2,933	2,77
390	test	130	1,5-hexadiene	Alcenos e alcinos	592-42-7	2,9	3,05
391	training	254	cis-2,trans-4-hexadiene	Alcenos e alcinos	5194-50-3	2,9	3,24
392	training	255	trans-2,trans-4-hexadiene	Alcenos e alcinos	5194-51-4	3,014	3,24
393	training	256	1-hexyne	Alcenos e alcinos	693-02-7	2,862	2,63
394	training	257	cyclohexanone	Compostos carbonílicos	108-94-1	1,818	1,03
395	test	131	5-hexen-2-one	Compostos carbonílicos	109-49-9	1,932	1,08
396	training	258	ethyl methacrylate	Éster	97-63-2	2,432	1,69
397	training	259	ethylacetoacetate	Éster	141-97-9	1,513	0,19
398	test	132	adipic acid	Ácido orgânico	124-04-9	1,421	0,13
399	test	133	diethyl oxalate	Éster	95-92-1	1,682	1,15
400	training	260	bromocyclohexane	Alcano halogenado	108-85-0	3,118	3,63
401	training	261	hexanenitrile	Nitrila	628-73-9	2,247	1,64
402	training	262	epsilon-caprolactam	Amida	105-60-2	1,274	-0,08
403	training	263	cyclohexanone oxime	Outros compostos	100-64-1	1,834	1,52
404	training	264	methylcyclopentane	Alcano	96-37-7	3,21	3,15
405	training	265	cyclohexane	Alcano	110-82-7	3,248	3,46
406	training	266	1-hexene	Alcenos e alcinos	592-41-6	3,227	3,38
407	test	134	4-methyl-1-pentene	Alcenos e alcinos	691-37-2	2,737	3,08
408	training	267	thiram	Organossulfurado	137-26-8	3,01	2,18
409	training	268	cyclohexanol	Álcool	108-93-0	2,046	1,35
410	training	269	hexanal	Compostos carbonílicos	66-25-1	2,345	2,37
411	test	135	2-hexanone	Compostos carbonílicos	591-78-6	2,128	1,45
412	training	270	3-methyl-2-pentanone	Compostos carbonílicos	565-61-7	0,63	1,48
413	test	136	4-methyl-2-pentanone	Compostos carbonílicos	108-10-1	0,63	1,31
414	training	271	hexanoic acid	Ácido orgânico	142-62-1	2,421	1,88
415	test	137	2-ethyl butyric acid	Ácido orgânico	88-09-5	2,291	1,74
416	training	272	butyl acetate	Éster	123-86-4	2,367	1,84
417	training	273	isobutyl acetate	Éster	110-19-0	2,345	1,74
418	training	274	sec-butyl acetate	Éster	105-46-4	2,313	1,97
419	training	275	tert-butyl acetate	Éster	540-88-5	2,334	1,88
420	test	138	hydroxycaproic acid	Ácido orgânico	1191-25-9	1,818	0,29
421	training	276	paraldehyde	Heterociclo	123-63-7	1,741	0,33
422	training	277	glucose	Outros compostos	50-99-7	-0,386	-2,57
423	training	278	1-bromohexane	Alcano halogenado	111-25-1	3,444	3,88
424	training	279	cyclohexylamine	Amina	108-91-8	2,188	1,30
425	test	139	hexane	Alcano	110-54-3	3,553	4,02
426	test	140	2,2-dimethylbutane	Alcano	75-83-2	3,455	3,74
427	test	141	2,3-dimethylbutane	Alcano	79-29-8	3,471	2,84
428	training	280	3-methylpentane	Alcano	96-14-0	3,335	3,98

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
429	training	281	lysine	Ácido orgânico	56-87-1	-0,282	-3,76
430	training	282	1-hexanol	Álcool	111-27-3	2,481	2,03
431	test	142	2-hexanol	Álcool	626-93-7	2,334	1,75
432	training	283	3-hexanol	Álcool	623-37-0	2,275	1,76
433	test	143	3,3-dimethyl-2-butanol	Álcool	464-07-3	2,182	1,75
434	test	144	dipropyl ether	Éter	111-43-3	2,481	2,04
435	test	145	diisopropyl ether	Éter	108-20-3	2,204	1,69
436	training	284	ethyl butyl ether	Éter	628-81-9	2,481	2,10
437	test	146	acetal	Éter	105-57-7	1,834	1,19
438	training	285	2-butoxyethanol	Álcool	111-76-2	1,829	0,78
439	training	286	dipropyl sulfone	Organossulfurado	598-03-8	1,589	0,36
440	training	287	diethylene glycol dimethyl ether	Éter	111-96-6	1,181	0,12
441	training	288	2-(2-ethoxyethoxy)ethanol	Álcool	111-90-0	1,083	-0,16
442	training	289	trimethylolpropane	Álcool	77-99-6	0,572	-0,76
443	test	147	sorbitol	Álcool	50-70-4	0,18	-2,68
444	test	148	hexylamine	Amina	111-26-2	2,498	1,98
445	training	290	di-propylamine	Amina	142-84-7	2,285	1,74
446	training	291	diisopropylamine	Amina	108-18-9	2,139	1,12
447	test	149	triethylamine	Amina	121-44-8	2,166	1,57
448	test	150	diisopropanolamine	Amina	110-97-4	0,931	-0,40
449	training	292	triethanolamine	Amina	102-71-6	0,833	-1,38
450	training	293	triethyl phosphate	Organofosforados	78-40-0	1,812	0,71
451	training	294	hexamethyl phosphoramide	Organofosforados	680-31-9	1,529	0,03
452	test	151	hexamethyldisiloxane	Outros compostos	107-46-0	3,662	2,89
453	test	152	3-nitrobenzotrifluoride	Nitrobenzeno	98-46-4	2,802	2,55
454	training	295	2-bromobenzoic acid	Ácido orgânico	88-65-3	2,574	2,54
455	training	296	3-bromobenzoic acid	Ácido orgânico	585-76-2	2,938	2,42
456	training	297	4-bromobenzoic acid	Ácido orgânico	586-76-5	2,933	2,43
457	training	298	o-chlorobenzoic acid	Ácido orgânico	118-91-2	2,492	2,39
458	training	299	chloramben	Ácido orgânico	133-90-4	1,25	2,05
459	test	153	4,5,6-trichloroguaiacol	Fenóis	2668-24-8	2,99	3,81
460	test	154	benzotrifluoride	Benzeno halogenado	98-08-8	3,036	2,91
461	test	155	benzonitrile	Derivados benzênicos	100-47-0	2,226	1,55
462	test	156	benzothiazole	Heterociclo aromático	95-16-9	2,47	2,13
463	test	157	2,4,6-trinitrotoluene	Nitrobenzeno	118-96-7	2,247	1,50
464	test	158	2,4-dichlorotoluene	Benzeno halogenado	95-73-8	3,684	3,95
465	training	300	3,4-dichlorophenyl urea	Fenil ureia	2327-02-8	2,49	2,35
466	training	301	2-(trifluoromethyl)aniline	Anilinas	88-17-5	2,36	2,24
467	test	159	3-(trifluoromethyl)aniline	Anilinas	98-16-8	2,36	2,23
468	training	302	4-(trifluoromethyl)aniline	Anilinas	455-14-1	2,36	2,30
469	training	303	3-(trifluoromethoxy)aniline	Anilinas	1535-73-5	2,36	2,37
470	test	160	2-(trifluoromethoxy)aniline	Anilinas	1535-75-7	2,36	2,45
471	training	304	4-(trifluoromethoxy)aniline	Anilinas	461-82-5	2,36	2,34
472	training	305	3-(trifluoromethylthio)aniline	Anilinas	369-68-6	2,36	2,88
473	test	161	4-(trifluoromethylthio)aniline	Anilinas	372-16-7	2,36	2,89
474	training	306	1H-benzimidazole	Heterociclo aromático	51-17-2	2,106	1,67
475	training	307	2-hydroxybenzimidazole	Outros compostos	615-16-7	1,986	0,74
476	training	308	2,4-dinitrotoluene	Nitrobenzeno	121-14-2	2,454	1,90
477	training	309	2,6-dinitrotoluene	Nitrobenzeno	606-20-2	2,519	1,81
478	test	162	3,4-dinitrotoluene	Nitrobenzeno	610-39-9	2,509	1,89
479	training	310	benzaldehyde	Derivados benzênicos	100-52-7	2,182	1,60
480	training	311	benzoic acid	Ácido orgânico	65-85-0	1,95	1,72
481	training	312	p-hydroxybenzaldehyde	Derivados benzênicos	123-08-0	2,111	1,27
482	training	313	salicylaldehyde	Derivados benzênicos	90-02-8	2,362	1,22
483	training	314	1,3-benzodioxole	Outros compostos	274-09-9	2,509	1,71
484	training	315	phenyl formate	Éster	1864-94-4	2,062	1,31
485	test	163	salicylic acid	Ácido orgânico	69-72-7	2,574	1,96
486	test	164	p-bromotoluene	Benzeno halogenado	106-38-7	3,237	3,35
487	test	165	(bromomethyl)benzene	Benzeno halogenado	100-39-0	2,965	2,76
488	training	316	(4-bromophenyl)urea	Fenil ureia	1967-25-5	2,12	2,10
489	training	317	benzyl chloride	Benzeno halogenado	100-44-7	2,628	2,51
490	training	318	o-chlorotoluene	Benzeno halogenado	95-49-8	3,237	3,27

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
491	test	166	p-chlorotoluene	Benzeno halogenado	106-43-4	3,189	3,30
492	training	319	2-chlorophenyl urea	Fenil ureia	114-38-5	1,61	1,74
493	test	167	3-chlorophenyl urea	Fenil ureia	1967-27-7	2,01	1,58
494	training	320	p-fluorotoluene	Benzeno halogenado	352-32-9	2,781	2,65
495	training	321	2-fluorophenyl urea	Fenil ureia	656-31-5	1,31	1,20
496	training	322	3-fluorophenyl urea	Fenil ureia	770-19-4	1,77	1,07
497	training	323	4-fluorophenyl urea	Fenil ureia	659-30-3	1,52	1,13
498	test	168	formanilide	Amida	103-70-8	2,003	1,20
499	test	169	m-nitrotoluene	Nitrobenzeno	99-08-1	2,71	2,32
500	test	170	o-nitrotoluene	Nitrobenzeno	88-72-2	2,628	2,32
501	training	324	p-nitrotoluene	Nitrobenzeno	99-99-0	2,693	2,34
502	test	171	o-nitroanisole	Nitrobenzeno	91-23-6	2,318	2,02
503	training	325	4-methyl-3-nitrophenol	Fenóis	2042-14-0	2,61	2,33
504	test	172	3-methyl-4-nitrophenol	Fenóis	2581-34-2	2,61	2,27
505	training	326	3-methyl-2-nitrophenol	Fenóis	4920-77-8	2,61	2,28
506	training	327	2-methyl-3-nitrophenol	Fenóis	5460-31-1	2,61	2,28
507	test	173	5-methyl-2-nitrophenol	Fenóis	700-38-9	2,61	2,29
508	training	328	toluene	Benzeno e Alquil benzeno	108-88-3	1,97	2,56
509	training	329	2-bromo-4-methylaniline	Anilinas	583-68-6	1,96	2,52
510	training	330	2-bromo-5-methylaniline	Anilinas	53078-85-6	1,96	2,51
511	test	174	3-bromo-4-methylaniline	Anilinas	7745-91-7	1,96	2,54
512	training	331	4-bromo-2-methylaniline	Anilinas	583-75-5	1,96	2,51
513	training	332	4-bromo-3-methylaniline	Anilinas	6933-10-4	1,96	2,55
514	test	175	5-bromo-2-methylaniline	Anilinas	39478-78-9	1,96	2,50
515	training	333	3-bromo-2-methylaniline	Anilinas	55289-36-6	1,96	2,50
516	training	334	3-chloroanisidine	Anilinas	5345-54-0	1,93	1,91
517	training	335	phenylurea	Fenil ureia	64-10-8	1,35	0,85
518	training	336	anisole	Derivados benzênicos	100-66-3	2,525	2,10
519	test	176	benzyl alcohol	Derivados benzênicos	100-51-6	1,948	1,07
520	test	177	m-cresol	Fenóis	108-39-4	1,54	1,93
521	test	178	o-cresol	Fenóis	95-48-7	1,34	1,89
522	training	337	p-cresol	Fenóis	106-44-5	1,69	1,95
524	test	179	guaiacol	Fenóis	90-05-1	1,6	1,32
525	test	180	p-methoxyphenol	Fenóis	150-76-5	1,75	1,31
526	training	338	3-methoxyphenol	Fenóis	150-19-6	1,55	1,32
527	test	181	benzylamine	Derivados benzênicos	100-46-9	1,97	0,90
528	test	182	N-methylaniline	Anilinas	100-61-8	2,28	1,68
529	training	339	m-toluidine	Anilinas	108-44-1	1,74	1,32
530	test	183	o-toluidine	Anilinas	95-53-4	1,74	1,32
531	training	340	p-toluidine	Anilinas	106-49-0	1,9	1,34
532	training	341	2,6-dimethylpyridine	Heterociclo aromático	108-48-5	2,291	1,60
533	test	184	m-toluenediamine	Derivados benzênicos	95-80-7	1,453	0,37
534	test	185	simazine	Heterociclo aromático	122-34-9	2,08	2,48
535	training	342	butyl acrylate	Éster	141-32-2	2,661	2,20
536	training	343	isobutyl acrylate	Éster	106-63-8	2,585	2,14
537	training	344	diethyl malonate	Éster	105-53-3	1,899	0,93
538	test	186	oxamyl	Organossulfurado	23135-22-0	0,9	-0,16
539	training	345	mevinphos	Organofosforados	7786-34-7	1,64	0,71
540	training	346	methylcyclohexane	Alcano	108-87-2	3,488	3,90
541	training	347	cycloheptane	Alcano	291-64-5	3,553	4,01
542	training	348	1-heptene	Alcenos e alcinos	592-76-7	3,548	4,00
543	training	349	aldicarb	Organossulfurado	116-06-3	1,3	1,58
544	test	187	2-heptanone	Compostos carbonílicos	110-43-0	2,454	1,92
545	training	350	5-methyl-2-hexanone	Compostos carbonílicos	110-12-3	2,4	1,88
546	training	351	2,4-dimethyl-3-pentanone	Compostos carbonílicos	565-80-0	2,389	1,91
547	training	352	cis-2-methylcyclohexanol	Álcool	7443-70-1	2,378	1,80
548	training	353	trans-2-methylcyclohexanol	Álcool	7443-52-9	2,367	1,80
549	test	188	heptanoic acid	Ácido orgânico	111-14-8	2,693	2,41
550	training	354	1-bromoheptane	Alcano halogenado	629-04-9	3,749	4,40
551	training	355	1-chloroheptane	Alcano halogenado	629-06-1	3,635	4,30
552	training	356	heptane	Alcano	142-82-5	3,825	4,33
553	training	357	1-heptanol	Álcool	111-70-6	2,802	2,53

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
554	training	358	2-heptanol	Álcool	543-49-7	2,634	2,34
555	training	359	3-heptanol	Álcool	589-82-2	2,596	2,29
556	test	189	4-heptanol	Álcool	589-55-9	2,585	2,26
557	training	360	heptylamine	Amina	111-68-2	2,775	2,57
558	test	190	phorate	Organofosforados	298-02-2	2,82	3,71
559	training	361	chlorothalonil	Derivados benzênicos	1897-45-6	2,98	3,98
560	training	362	phthalic anhydride	Derivados benzênicos	85-44-9	2,247	0,89
561	training	363	ethynylbenzene	Derivados benzênicos	536-74-3	2,683	2,50
562	training	364	3,6-dichloro-2-methoxybenzoic acid	Ácido orgânico	1918-00-9	0,99	2,65
563	training	365	(2,4-dichlorophenoxy)acetic acid	Ácido orgânico	94-75-7	2,11	2,82
564	training	366	quinoxaline	Heterociclo aromático	91-19-0	1,965	1,12
565	test	191	benzofuran	Heterociclo poliaromático	271-89-6	2,829	2,75
566	training	367	isophthalic acid	Ácido orgânico	121-91-5	2,28	1,04
567	training	368	phthalic acid	Ácido orgânico	88-99-3	1,774	1,22
568	training	369	terephthalic acid	Ácido orgânico	100-21-0	2,465	1,01
569	training	370	benzothiophene	Heterociclo poliaromático	95-15-8	3,074	3,24
570	training	371	3-(trifluoromethylphenyl) urea	Fenil ureia	13114-87-9	1,6	1,73
571	training	372	indole	Heterociclo aromático	120-72-9	2,541	2,29
572	training	373	benzeneacetonitrile	Derivados benzênicos	140-29-4	2,226	1,42
573	training	374	styrene	Derivados benzênicos	100-42-5	3,036	2,92
574	training	375	1,3,5,7-cyclooctatetraene	Alcenos e alcinos	629-20-9	3,053	3,10
575	training	376	acetophenone	Derivados benzênicos	98-86-2	2,264	1,65
576	test	192	benzeneacetaldehyde	Derivados benzênicos	122-78-1	2,345	1,75
577	training	377	2-methylbenzaldehyde	Derivados benzênicos	529-20-4	2,606	1,91
578	training	378	2,3-dihydrobenzofuran	Derivados benzênicos	496-16-2	2,541	2,16
579	training	379	phenyloxirane	Derivados benzênicos	96-09-3	2,253	1,72
580	training	380	methyl benzoate	Éster	93-58-3	2,574	1,98
581	training	381	o-toluic acid	Ácido orgânico	118-90-1	2,639	2,03
582	training	382	p-toluic acid	Ácido orgânico	99-94-5	2,65	2,12
583	training	383	benzeneacetic acid	Ácido orgânico	103-82-2	2,144	1,72
584	training	384	phenyl acetate	Éster	122-79-2	2,188	1,59
585	training	385	m-toluic acid	Ácido orgânico	99-04-7	2,666	2,08
586	test	193	methyl salicylate	Éster	119-36-8	2,764	2,07
587	training	386	vanillin	Derivados benzênicos	121-33-5	2,035	1,31
588	test	194	acetanilide	Amida	103-84-4	2,008	1,05
589	test	195	ethylbenzene	Benzeno e Alquil benzeno	100-41-4	2,73	3,27
590	training	387	o-xylene	Benzeno e Alquil benzeno	95-47-6	2,7	3,16
591	test	196	m-xylene	Benzeno e Alquil benzeno	108-38-3	2,46	3,15
592	training	388	p-xylene	Benzeno e Alquil benzeno	106-42-3	2,77	3,15
594	training	389	methyl parathion	Organofosforados	298-00-0	2,64	2,97
596	training	390	phenetole	Derivados benzênicos	103-73-1	2,742	2,56
597	training	391	2-phenylethanol	Derivados benzênicos	60-12-8	2,117	1,51
598	training	392	3-methylbenzenemethanol	Derivados benzênicos	587-03-1	2,247	1,53
599	training	393	4-methylbenzenemethanol	Derivados benzênicos	589-18-4	2,237	1,54
600	training	394	1-phenylethanol	Derivados benzênicos	98-85-1	1,57	1,58
601	test	197	o-ethylphenol	Fenóis	90-00-6	2,721	2,45
602	training	395	m-ethylphenol	Fenóis	620-17-7	2,737	2,53
603	training	396	p-ethylphenol	Fenóis	123-07-9	2,737	2,54
604	test	198	2,3-xylenol	Fenóis	526-75-0	2,66	2,34
605	training	397	2,4-xylenol	Fenóis	105-67-9	2,66	2,37
606	training	398	2,5-xylenol	Fenóis	95-87-4	2,66	2,35
607	test	199	2,6-xylenol	Fenóis	576-26-1	2,66	2,32
608	training	399	3,4-xylenol	Fenóis	95-65-8	2,66	2,41
609	training	400	3,5-xylenol	Fenóis	108-68-9	2,66	2,38
610	training	401	benzyl methyl ether	Derivados benzênicos	538-86-3	2,111	1,69
611	training	402	2-methylanisole	Derivados benzênicos	578-58-5	2,868	2,60
612	training	403	3-methylanisole	Derivados benzênicos	100-84-5	2,824	2,63
613	training	404	4-methylanisole	Derivados benzênicos	104-93-8	2,906	2,63
614	training	405	1,2-dimethoxybenzene	Derivados benzênicos	91-16-7	2,509	2,10
615	test	200	endothall	Ácido orgânico	145-73-3	2,14	0,55
616	test	201	N,N-dimethylaniline	Anilinas	121-69-7	2,634	2,05
617	training	406	o-ethylaniline	Anilinas	578-54-1	2,324	1,91

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
618	test	202	2,4,6-trimethylpyridine	Heterociclo aromático	108-75-8	2,4	2,10
619	test	203	benzeneethanamine	Derivados benzênicos	64-04-0	2,144	1,41
620	training	407	1,5-cyclooctadiene	Alcenos e alcinos	111-78-4	3,096	3,58
621	training	408	vinylcyclohexene	Alcenos e alcinos	100-40-3	3,515	3,49
622	test	204	2,5-dimethyl-2,4-hexadiene	Alcenos e alcinos	764-13-6	3,281	3,55
623	training	409	butyl methacrylate	Éster	97-88-1	2,944	2,59
624	test	205	diethyl succinate	Éster	123-25-1	2,03	1,25
625	training	410	octanenitrile	Nitrila	124-12-9	2,873	2,76
626	test	206	cyclooctane	Alcano	292-64-8	3,798	4,62
627	training	411	1-octene	Alcenos e alcinos	111-66-0	3,863	4,61
628	test	207	2,4,4-trimethyl-1-pentene	Alcenos e alcinos	107-39-1	3,852	4,03
629	training	412	2-octanone	Compostos carbonílicos	111-13-7	2,666	2,54
630	training	413	octanoic acid	Ácido orgânico	124-07-2	3,036	2,92
631	test	208	1-bromooctane	Alcano	111-83-1	4,037	4,91
632	test	209	octane	Alcano	111-65-9	4,179	4,73
633	training	414	1-octanol	Álcool	111-87-5	3,047	3,21
634	test	210	2-octanol	Álcool	123-96-6	2,955	2,96
635	training	415	4-octanol	Álcool	589-62-8	2,835	2,83
636	training	416	dibutyl ether	Éter	142-96-1	3,123	3,04
637	training	417	diethylene glycol diethyl ether	Éter	112-36-7	1,589	0,64
638	training	418	diethylene glycol monobutyl ether	Éter	112-34-5	1,682	0,63
639	training	419	octylamine	Amina	111-86-4	2,955	3,24
640	training	420	dibutylamine	Amina	111-92-2	2,917	2,71
641	training	421	octamethylcyclotetrasiloxane	Outros compostos	556-67-2	4,151	3,56
642	test	211	folpet	Organossulfurado	133-07-3	3,27	2,92
643	training	422	2H-1-benzopyran-2-one	Derivados benzênicos	91-64-5	2,133	1,72
644	training	423	1H-indene-1,3(2H)-dione	Derivados benzênicos	606-23-5	1,709	1,54
645	test	212	isoquinoline	Heterociclo poliaromático	119-65-3	2,509	2,14
646	training	424	quinoline	Heterociclo poliaromático	91-22-5	2,481	2,19
647	training	425	cinnamonitrile	Derivados benzênicos	4360-47-8	2,443	2,01
648	training	426	8-hydroxyquinoline	Heterociclo poliaromático	148-24-3	2,476	1,91
649	training	427	indene	Derivados benzênicos	95-13-6	2,965	3,04
650	test	213	captan	Organossulfurado	133-06-2	2,3	3,00
651	training	428	2-methylbenzofuran	Heterociclo aromático	4265-25-2	3,129	3,07
652	training	429	2-propenophenone	Derivados benzênicos	768-03-6	2,4	1,82
653	test	214	cinnamic acid	Ácido orgânico	621-82-9	2,536	2,38
654	training	430	(4-chloro-2-methylphenoxy)acetic acid	Ácido orgânico	94-74-6	3,86	2,41
655	training	431	propanil	Amida	709-98-8	2,48	3,04
656	training	432	benzenepropanenitrile	Derivados benzênicos	645-59-0	2,313	1,94
657	training	433	cinnamamide	Amida	621-79-4	2,144	1,19
658	training	434	indane	Derivados benzênicos	496-11-7	3,189	2,97
659	training	435	α-methylstyrene	Derivados benzênicos	98-83-9	3,27	3,31
660	training	436	chlorbromuron	Fenil ureia	13360-45-7	2,58	3,02
661	test	215	imidacloprid	Heterociclo aromático	105827-78-9	2,64	0,65
662	training	437	diuron	Fenil ureia	330-54-1	2,82	2,92
663	training	438	linuron	Fenil ureia	330-55-2	2,43	2,82
664	training	439	2,3-dihydro-1H-inden-1-ol	Derivados benzênicos	6351-10-6	4,06	1,59
665	test	216	2,3-dihydro-1H-inden-5-ol	Fenóis	1470-94-6	4,06	2,37
666	test	217	4-methylacetophenone	Derivados benzênicos	122-00-9	2,568	2,11
667	training	440	1-phenyl-1-propanone	Derivados benzênicos	93-55-0	2,568	2,15
668	training	441	1-phenyl-2-propanone	Derivados benzênicos	103-79-7	2,16	1,70
669	training	442	ethyl benzoate	Éster	93-89-0	2,813	2,39
670	test	218	benzyl acetate	Éster	140-11-4	2,443	2,07
671	training	443	4-methylphenyl acetate	Éster	140-39-6	2,525	1,96
672	training	444	(±)-2-phenylpropionic acid	Ácido orgânico	492-37-5	2,356	2,17
673	training	445	ethyl vanillin	Derivados benzênicos	121-32-4	2,237	1,82
674	test	219	metobromuron	Fenil ureia	3060-89-7	2,02	2,18
675	training	446	monuron	Fenil ureia	150-68-5	1,7	1,96
676	training	447	3-(4-chlorophenyl)-1-methoxy-1-methylurea	Fenil ureia	1746-81-2	1,84	1,99
677	training	448	chlorpyrifos	Organofosforados	2921-88-2	3,79	5,15
678	training	449	1,1-dimethyl-3-(3-fluorophenyl) urea	Fenil ureia	330-39-2	1,73	1,32
679	training	450	1,1-dimethyl-3-(4-fluorophenyl) urea	Fenil ureia	332-33-2	1,43	1,46

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
680	training	451	1,2,3,4-tetrahydroquinoline	Derivados benzênicos	635-46-1	2,623	2,27
681	test	220	p-dimethylaminobenzaldehyde	Derivados benzênicos	100-10-7	2,362	1,80
682	training	452	cumene	Benzeno e Alquil benzeno	98-82-8	3,368	3,67
683	training	453	m-ethyltoluene	Benzeno e Alquil benzeno	620-14-4	3,542	3,79
684	training	454	o-ethyltoluene	Benzeno e Alquil benzeno	611-14-3	3,297	3,87
685	training	455	p-ethyltoluene	Benzeno e Alquil benzeno	622-96-8	3,352	3,83
686	test	221	1,2,3-trimethylbenzene	Benzeno e Alquil benzeno	526-73-8	3,335	3,63
687	training	456	1,2,4-trimethylbenzene	Benzeno e Alquil benzeno	95-63-6	3,352	3,62
688	training	457	mesitylene	Benzeno e Alquil benzeno	108-67-8	3,237	3,64
689	test	222	propylbenzene	Benzeno e Alquil benzeno	103-65-1	3,384	3,86
690	training	458	fenitrothion	Organofosforados	122-14-5	3,51	3,31
691	training	459	benzyl ethyl ether	Derivados benzênicos	539-30-0	2,552	2,38
692	training	460	benzenepropanol	Derivados benzênicos	122-97-4	2,4	2,00
693	test	223	2-propylphenol	Fenóis	644-35-9	2,971	2,95
694	training	461	4-propylphenol	Fenóis	645-56-7	3,118	3,01
695	test	224	2,3,4-trimethylphenol	Fenóis	526-85-2	3,76	2,75
696	training	462	2,3,5-trimethylphenol	Fenóis	697-82-5	3,76	2,73
697	training	463	2,3,6-trimethylphenol	Fenóis	2416-94-6	3,76	2,72
698	test	225	2,4,5-trimethylphenol	Fenóis	496-78-6	3,76	2,75
699	training	464	2,4,6-trimethylphenol	Fenóis	527-60-6	3,76	2,72
700	training	465	3,4,5-trimethylphenol	Fenóis	527-54-8	3,76	2,77
701	training	466	bromacil	Heterociclo aromático	314-40-9	1,97	1,20
702	training	467	terbacil	Heterociclo aromático	5902-51-2	1,63	1,78
703	test	226	cyanazine	Triazinas	21725-46-2	2,26	2,05
704	training	468	amphetamine	Derivados benzênicos	300-62-9	2,334	1,85
705	training	469	N,N-dimethylbenzylamine	Derivados benzênicos	103-83-3	2,454	1,84
706	training	470	isophorone	Compostos carbonílicos	78-59-1	2,302	1,90
707	test	227	glyceryl triacetate	Éster	102-76-1	1,513	0,40
708	test	228	propazine	Triazinas	139-40-2	2,19	2,94
709	test	229	triethazine	Triazinas	1912-26-1	2,74	3,58
710	test	230	azelaic acid	Ácido orgânico	123-99-9	2,231	1,37
711	training	471	ametryn	Triazinas	834-12-8	2,13	3,09
712	training	472	1-nonene	Alcenos e alcinos	124-11-8	4,179	5,14
713	test	231	2-nonanone	Compostos carbonílicos	821-55-6	3,096	3,08
714	training	473	5-methyl-2-octanone	Compostos carbonílicos	58654-67-4	2,965	3,07
715	test	232	nonanoic acid	Ácido orgânico	112-05-0	3,237	3,47
716	test	233	nonane	Alcano	111-84-2	4,451	5,24
717	training	474	1-nonanol	Álcool	143-08-8	3,564	3,76
718	test	234	2,6-dimethyl-4-heptanol	Álcool	108-82-7	3,053	3,03
719	training	475	tripropylamine	Amina	102-69-2	2,895	3,08
720	training	476	terbufos	Organofosforados	13071-79-9	2,5	4,61
721	training	477	ethion	Organofosforados	563-12-2	3,94	4,74
722	training	478	chlordane	Outros compostos	57-74-9	5,15	6,02
724	training	479	a-chlordan	Outros compostos	5103-71-9	5,15	6,02
725	test	235	trans-chlordane	Outros compostos	5103-74-2	5,15	6,02
726	training	480	2-hydroxy-1,4-naphthalenedione	Derivados benzênicos	83-72-7	2,171	0,99
727	test	236	1-chloronaphthalene	HPA	90-13-1	3,499	3,95
728	test	237	2-chloronaphthalene	HPA	91-58-7	3,542	3,91
729	training	481	naphthalene	HPA	91-20-3	3	3,33
730	test	238	azulene	HPA	275-51-4	3,129	3,51
731	training	482	1-naphthol	HPA	90-15-3	2,922	2,79
732	test	239	2-naphthol	HPA	135-19-3	2,846	2,93
733	training	483	captafol	Outros compostos	2425-06-1	3,32	3,57
734	test	240	2-methylquinoline	Heterociclo poliaromático	91-63-4	2,786	2,66
735	test	241	1-naphthylamine	HPA	134-32-7	3,58	2,27
736	training	484	2-naphthylamine	HPA	91-59-8	3,58	2,30
737	test	242	metamitron	Heterociclo aromático	41394-05-2	1,55	1,17
738	training	485	benzalacetone	Derivados benzênicos	122-57-6	2,503	2,23
739	training	486	methyl cinnamate	Éster	1754-62-7	2,802	2,58
740	test	243	dimethyl phthalate	Éster	131-11-3	1,63	1,96
741	test	244	dimethyl terephthalate	Éster	120-61-6	2,601	1,83
742	training	487	Fluometuron	Fenil ureia	2164-17-2	1,96	2,16

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
743	training	488	1,2,3,4-tetrahydronaphthalene	Derivados benzênicos	119-64-2	3,276	3,79
744	training	489	3-phenyl-1-cyclopropyl urea	Fenil ureia	13140-86-8	1,72	1,62
745	training	490	azinphos-methyl	Organofosforados	86-50-0	2,69	2,75
746	training	491	isopropyl benzoate	Éster	939-48-0	3,107	2,72
747	test	245	Chlorotoluron	Fenil ureia	15545-48-9	2,43	2,25
749	training	492	isopropyl phenylcarbamate	Derivados benzênicos	122-42-9	1,95	2,60
750	training	493	butylbenzene	Benzeno e Alquil benzeno	104-51-8	3,694	4,34
751	training	494	isobutylbenzene	Benzeno e Alquil benzeno	538-93-2	3,558	4,13
752	test	246	sec-butylbenzene	Benzeno e Alquil benzeno	135-98-8	3,863	4,36
753	training	495	tert-butylbenzene	Benzeno e Alquil benzeno	98-06-6	3,613	4,49
754	training	496	m-cymene	Benzeno e Alquil benzeno	535-77-3	3,825	4,07
755	test	247	o-cymene	Benzeno e Alquil benzeno	527-84-4	3,76	4,11
756	training	497	p-cymene	Benzeno e Alquil benzeno	99-87-6	3,607	4,17
757	training	498	o-diethylbenzene	Benzeno e Alquil benzeno	135-01-3	3,781	4,55
758	training	499	m-diethylbenzene	Benzeno e Alquil benzeno	141-93-5	3,863	4,38
759	test	248	p-diethylbenzene	Benzeno e Alquil benzeno	105-05-5	3,869	4,36
760	test	249	3-ethyl-o-xylene	Benzeno e Alquil benzeno	933-98-2	3,738	4,40
761	test	250	4-ethyl-o-xylene	Benzeno e Alquil benzeno	934-80-5	3,825	4,33
762	training	500	2-ethyl-m-xylene	Benzeno e Alquil benzeno	2870-04-4	3,705	4,40
763	training	501	4-ethyl-m-xylene	Benzeno e Alquil benzeno	874-41-9	3,809	4,32
764	training	502	5-ethyl-m-xylene	Benzeno e Alquil benzeno	934-74-7	3,852	4,34
765	test	251	2-ethyl-p-xylene	Benzeno e Alquil benzeno	1758-88-9	3,787	4,35
766	training	503	1,2,3,4-tetramethylbenzene	Benzeno e Alquil benzeno	488-23-3	3,553	4,07
767	training	504	1,2,3,5-tetramethylbenzene	Benzeno e Alquil benzeno	527-53-7	3,607	4,06
768	test	252	1,2,4,5-tetramethylbenzene	Benzeno e Alquil benzeno	95-93-2	3,607	4,05
769	training	505	parathion	Organofosforados	56-38-2	3,02	3,76
770	training	506	1,1-dimethyl-3-(4-methoxy-phenyl) urea	Fenil ureia	28170-54-9	1,72	1,18
771	test	253	4-butylphenol	Fenóis	1638-22-8	3,363	3,55
772	test	254	N,N-dietyhlaniline	Anilinas	91-66-7	3,178	3,43
773	training	507	α-pinene	Alcenos e alcinos	80-56-8	4,005	3,66
774	training	508	γ-terpinene	Alcenos e alcinos	99-85-4	3,825	4,36
775	test	255	terpinolene	Alcenos e alcinos	586-62-9	3,809	3,82
776	training	509	ipazine	Triazinas	1912-25-0	3,39	3,65
777	training	510	methoxypropazine	Triazinas	1610-18-0	2,43	2,80
778	training	511	prometryn	Triazinas	7287-19-6	2,8	3,31
779	training	512	2-decanone	Compostos carbonílicos	693-54-9	3,428	3,63
780	training	513	decanoic acid	Ácido orgânico	334-48-5	3,602	3,93
781	test	256	decane	Alcano	124-18-5	4,777	5,87
782	training	514	2,2,3,3-tetramethylhexane	Alcano	13475-81-5	4,113	5,64
783	training	515	1-decanol	Álcool	112-30-1	3,863	4,24
784	test	257	2-(2-furyl)benzimidazole	Heterociclo poliaromático	3878-19-1	2,55	2,84
785	training	516	2-methyl-1,4-naphthalenedione	Derivados benzênicos	58-27-5	2,574	1,91
786	training	517	2-hydroxy-3-methyl-1,4-naphthalenedione	Derivados benzênicos	483-55-6	2,03	1,48
787	test	258	2-methoxy-1,4-naphthalenedione	Derivados benzênicos	2348-82-5	2,111	1,70
788	training	518	4-phenylpyridine	Heterociclo aromático	939-23-1	2,786	2,40
789	training	519	1-methylnaphthalene	HPA	90-12-0	3,482	3,84
790	test	259	2-methylnaphthalene	HPA	91-57-6	3,553	3,83
791	test	260	1-naphthalenemethanol	HPA	4780-79-4	2,22	2,17
792	training	520	2-naphthalenemethanol	HPA	1592-38-7	2,22	2,28
793	training	521	chloramphenicol	Derivados benzênicos	56-75-7	1,997	1,15
794	training	522	butyl benzoate	Éster	136-60-7	3,466	3,40
795	test	261	methiocarb	Derivados benzênicos	2032-65-7	2,25	2,54
796	training	523	pentylbenzene	Benzeno e Alquil benzeno	538-68-1	4,043	4,81
797	training	524	pentamethylbenzene	Benzeno e Alquil benzeno	700-12-9	3,858	4,42
798	test	262	2-undecanone	Compostos carbonílicos	112-12-9	3,602	4,25
799	test	263	methyl decanoate	Éster	110-42-9	3,776	4,58
800	test	264	1-undecanol	Álcool	112-42-5	3,945	4,83
801	test	265	decachlorobiphenyl	Bifenil	2051-24-3	5,87	8,59
802	training	525	2,2′,3,3′,4,5,5′,6,6′ nonachlorobiphenyl	Bifenil	52663-77-1	5,816	8,34
803	training	526	2,2′,3,3′,5,5′,6,6′ -octachlorobiphenyl	Bifenil	2136-99-4	5,239	8,07
804	training	527	2,2′,3,3′,4,4′,6-heptachlorobiphenyl	Bifenil	52663-71-5	5,022	7,69
805	test	266	fiponil	Derivados benzênicos	120068-37-3	3,08	4,40

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
806	training	528	2,2′,3,3′,4,4′ hexachlorobiphenyl	Bifenil	38380-07-3	3,83	7,27
807	test	267	2,2′,4,4′,6,6′ -hexachlorobiphenyl	Bifenil	33979-03-2	3,83	7,27
808	training	529	2,2′,3,3′,6,6′ -hexachlorobiphenyl	Bifenil	38411-22-2	3,83	7,28
809	training	530	2,3,3′,4,4′,5-hexachlorobiphenyl	Bifenil	38380-08-4	3,83	7,26
810	test	268	2,2′,4,4′,5′ 5′ -hexachloro-1,1′ -biphenyl	Bifenil	35065-27-1	3,83	7,25
811	training	531	3,3′,4,4′,5,5′ -hexachlorobiphenyl	Bifenil	32774-16-6	3,83	7,26
812	training	532	2,3,4,5,6-pentachlorobiphenyl	Bifenil	18259-05-7	4,804	6,79
813	training	533	2,2′,4,5,5′ -pentachlorobiphenyl	Bifenil	37680-73-2	4,859	6,77
815	training	534	2,3,4,5-tetrachlorobiphenyl	Bifenil	33284-53-6	5,64	6,14
816	training	535	2,2′,4′,5-tetrachlorobiphenyl	Bifenil	41464-40-8	5,64	6,23
818	training	536	3,3′,4,4′ -tetrachlorobiphenyl	Bifenil	32598-13-3	5	6,21
819	test	269	2,2′,3,3′ -tetrachlorobiphenyl	Bifenil	38444-93-8	5	6,21
820	training	537	2,2′,5,5′ -tetrachlorobiphenyl	Bifenil	35693-99-3	5,37	6,24
821	test	270	2,2′,6,6′ -tetrachlorobiphenyl	Bifenil	15968-05-5	4,89	6,22
822	test	271	2,3′,4′,5-tetrachlorobiphenyl	Bifenil	32598-11-1	4,85	6,22
823	training	538	2,4,5-trichlorobiphenyl	Bifenil	15862-07-4	5,21	5,69
824	test	272	2,4,6-trichlorobiphenyl	Bifenil	35693-92-6	5,21	5,70
825	training	539	2,2′,5-trichlorobiphenyl	Bifenil	37680-65-2	5,21	5,71
826	training	540	2,3′,4′ -trichlorobiphenyl	Bifenil	38444-86-9	5,21	5,69
827	test	273	2,3,5-trichlorobiphenyl	Bifenil	38444-81-4	5,21	5,70
828	training	541	2,4,4′-trichlorobiphenyl	Bifenil	7012-37-5	5,21	5,70
829	training	542	2,4′,5-trichlorobiphenyl	Bifenil	16606-02-3	5,21	5,70
831	test	274	2,5-dichlorobiphenyl	Bifenil	34883-39-1	4,7	5,14
832	training	543	2,6-dichlorobiphenyl	Bifenil	33146-45-1	4,7	5,15
833	training	544	3,3′ -dichloro-1,1′ -biphenyl	Bifenil	2050-67-1	4,7	5,13
834	test	275	4,4′ -dichloro-1,1′ -biphenyl	Bifenil	2050-68-2	5,27	5,12
835	test	276	2,2′ -dichloro-1,1′ -biphenyl	Bifenil	13029-08-8	4,7	5,15
836	training	545	3,4-dichloro-1,1′-biphenyl	Bifenil	2974-92-7	4,7	5,13
838	training	546	2,4′ -dichloro-1,1′ -biphenyl	Bifenil	34883-43-7	4,55	5,14
839	training	547	dieldrin	Alceno halogenado	60-57-1	4,11	4,98
840	test	277	endrin	Alceno halogenado	72-20-8	4,2	4,98
841	test	278	1,10-phenanthroline	Heterociclo poliaromático	66-71-7	2,373	2,31
842	test	279	phenazine	Heterociclo poliaromático	92-82-0	2,922	2,82
843	test	280	dibenzofuran	Heterociclo poliaromático	132-64-9	3,618	3,92
844	training	548	2-chlorobiphenyl	Bifenil	2051-60-7	3,836	4,59
845	training	549	3-chlorobiphenyl	Bifenil	2051-61-8	3,869	4,57
846	training	550	4-chlorobiphenyl	Bifenil	2051-62-9	3,885	4,57
847	test	281	norflurazon	Heterociclo aromático	27314-13-2	2,66	2,66
848	training	551	dibenzopyrrole	Heterociclo poliaromático	86-74-8	3,401	3,69
849	test	282	acenaphthene	HPA	83-32-9	3,531	4,01
850	test	283	phenylbenzene	Bifenil	92-52-4	3,04	4,02
852	training	552	azobenzene	Derivados benzênicos	17082-12-1	3,13	4,30
853	training	553	diphenyl ether	Derivados benzênicos	101-84-8	3,667	3,68
854	training	554	diphenyl sulfide	Derivados benzênicos	139-66-2	3,798	4,36
855	test	284	p-aminodiphenyl	Bifenil	92-67-1	2,933	2,89
856	test	285	diphenylamine	Anilinas	122-39-4	2,78	3,34
857	training	555	carbaryl	Derivados benzênicos	63-25-2	2,02	2,45
858	training	556	p-aminoazobenzene	Derivados benzênicos	60-09-3	3,232	4,02
859	training	557	1,2-dimethylnaphthalene	HPA	573-98-8	3,722	4,38
860	training	558	1,3-dimethylnaphthalene	HPA	575-41-7	3,781	4,36
861	training	559	1,4-dimethylnaphthalene	HPA	571-58-4	3,754	4,37
862	training	560	1,5-dimethylnaphthalene	HPA	571-61-9	3,76	4,37
863	training	561	1,7-dimethylnaphthalene	HPA	575-37-1	3,792	4,36
864	training	562	2,3-dimethylnaphthalene	HPA	581-40-8	3,771	4,37
865	training	563	2,6-dimethylnaphthalene	HPA	581-42-0	3,722	4,37
866	training	564	1-ethylnaphthalene	HPA	1127-76-0	3,771	4,47
867	training	565	2-ethylnaphthalene	HPA	939-27-5	3,76	4,47
868	training	566	hydrazobenzene	Derivados benzênicos	122-66-7	2,976	2,88
869	training	567	p-benzidine	Bifenil	92-87-5	5,36	1,59
870	training	568	chlorfenvinphos	Organofosforados	470-90-6	2,47	4,05
871	training	569	4-phenylcyclohexanone	Compostos carbonílicos	4894-75-1	2,71	2,62
872	training	570	diethyl phthalate	Éster	84-66-2	2,721	2,60

Mol ID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
873	training	571	carbofuran	Derivados benzênicos	1563-66-2	1,79	2,08
874	training	572	hexylbenzene	Benzeno e Alquil benzeno	1077-16-3	4,38	5,27
875	test	286	hexamethylbenzene	Benzeno e Alquil benzeno	87-85-4	3,928	4,71
876	training	573	isoproturon	Fenil ureia	34123-59-6	2,35	2,63
877	training	574	diazinon	Organofosforados	333-41-5	2,36	4,45
878	training	575	cyclododecanone	Compostos carbonílicos	830-13-7	3,607	4,45
879	test	287	dodecanoic acid	Ácido orgânico	143-07-7	3,879	5,13
880	training	576	dodecane	Alcano	112-40-3	4,695	6,42
881	training	577	1-dodecanol	Álcool	112-53-8	4,168	5,36
882	test	288	diethylene glycol dibutyl ether	Éter	112-73-2	2,421	2,48
883	test	289	hexachlorophene	Derivados benzênicos	70-30-4	3,515	6,77
884	training	578	9H-fluoren-9-one	HPA	486-25-9	3,325	3,45
885	training	579	acridine	Heterociclo poliaromático	260-94-6	3,227	3,51
886	training	580	fluorene	HPA	86-73-7	3,662	4,26
887	training	581	mantuamycin	Organofosforados	21609-90-5	5,07	6,37
888	training	582	benzophenone	Derivados benzênicos	119-61-9	2,64	3,03
889	test	290	phenyl benzoate	Éster	93-99-2	3,33	3,38
890	test	291	N-phenylbenzamide	Amida	93-98-1	2,802	2,43
891	training	583	diphenylmethane	Derivados benzênicos	101-81-5	3,629	4,33
892	training	584	4-methylbiphenyl	Bifenil	644-08-6	3,896	4,42
893	training	585	4-phenoxyphenyl urea	Fenil ureia	78508-44-8	2,56	2,49
894	training	586	benzyl phenyl ether	Derivados benzênicos	946-80-5	3,439	3,63
895	training	587	diphenylmethanol	Derivados benzênicos	91-01-0	2,829	2,76
896	test	292	4-biphenylmethanol	Bifenil	3597-91-9	2,69	3,18
897	training	588	imazapyr	Ácido orgânico	81334-34-1	2,35	1,52
899	training	589	trifluralin	Anilinas	1582-09-8	4,49	5,09
900	training	590	3-phenyl-1-cyclohexyl urea	Fenil ureia	886-59-9	2,07	3,25
901	test	293	fenamiphos	Organofosforados	22224-92-6	2,52	3,05
902	training	591	1-tridecanol	Álcool	112-70-9	4,543	5,71
903	test	294	anthraquinone	Derivados benzênicos	84-65-1	3,221	2,83
904	test	295	anthracene	HPA	120-12-7	3,858	4,56
905	test	296	diphenylacetylene	Derivados benzênicos	501-65-5	3,977	4,17
906	test	297	phenanthrene	HPA	85-01-8	3,77	4,55
907	test	298	2-anthracenamine	HPA	613-13-8	4,48	3,69
908	training	592	trans-stilbene	Derivados benzênicos	103-30-0	3,994	4,58
909	training	593	1-methylfluorene	HPA	1730-37-6	4,081	4,56
910	training	594	2-phenylacetophenone	Derivados benzênicos	451-40-1	3,107	3,15
911	training	595	benzyl benzoate	Éster	120-51-4	3,537	3,43
912	training	596	1,2-diphenylethane	Derivados benzênicos	103-29-7	3,934	4,74
913	training	597	4,4′-dimethylbiphenyl	Bifenil	613-33-2	4,146	4,97
914	training	598	dibenzyl ether	Derivados benzênicos	103-50-4	3,178	3,42
915	test	299	triadimenol	Outros compostos	55219-65-3	1,95	2,88
916	test	300	alachlor	Derivados benzênicos	15972-60-8	2,48	3,02
917	training	599	octylbenzene	Benzeno e Alquil benzeno	2189-60-8	4,804	6,46
918	training	600	tetradecanoic acid	Ácido orgânico	544-63-8	4,695	6,10
919	test	301	tetradecane	Alcano	629-59-4	5,294	7,70
920	training	601	1-tetradecanol	Álcool	112-72-1	4,657	6,21
921	training	602	9-anthracenecarboxylic acid	Ácido orgânico	723-62-6	2,54	3,84
922	training	603	2-phenyl-1H-indene-1,3(2H)-dione	Derivados benzênicos	83-12-5	2,955	3,10
923	training	604	2-methylanthracene	HPA	613-12-7	4,179	5,06
924	training	605	9-methylanthracene	HPA	779-02-2	4,135	5,06
925	training	606	1-methylphenanthrene	HPA	832-69-9	4,173	5,05
926	test	302	9-anthracenemethanol	HPA	1468-95-7	3,43	3,55
927	test	303	prochloraz	Heterociclo aromático	67747-09-5	4,13	3,78
928	training	607	bisphenol a	Fenóis	80-05-7	3,183	3,81
929	training	608	imazamethabenz	Outros compostos	100728-84-5	2,04	2,26
930	training	609	nicosulfuron	Outros compostos	111991-09-4	2,26	0,59

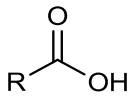
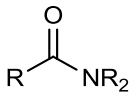
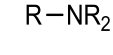
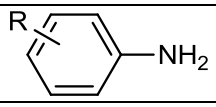
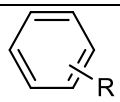
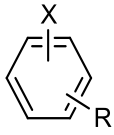
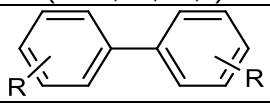
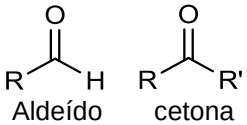
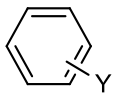
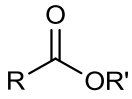
Conclusão

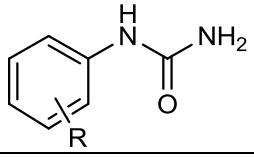
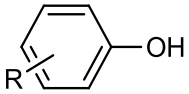
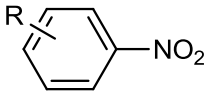
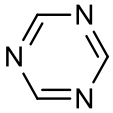
MolID	Status	Ord. Status	Nome	Classe	CAS	Exp logKoc	ALOGPs
931	training	610	metalaxyl	Derivados benzênicos	57837-19-1	1,66	1,47
932	training	611	metolachlor	Derivados benzênicos	51218-45-2	2,2	3,37
933	training	612	nonylbenzene	Benzeno e Alquil benzeno	1081-77-2	5,245	7,00
934	test	304	2,6-di-tert-butyl-p-cresol	Fenóis	128-37-0	4,151	5,25
935	training	613	Fluoranthene	HPA	206-44-0	4,135	5,04
936	test	305	pyrene	HPA	129-00-0	4,8	5,19
937	training	614	9,10-dimethylphenanthrene	HPA	604-83-1	4,472	5,40
938	training	615	dibutyl phthalate	Éster	84-74-2	3,945	4,53
939	test	306	tebuconazole	Outros compostos	107534-96-3	2,67	3,60
940	test	307	decylbenzene	Benzeno e Alquil benzeno	104-72-3	5,375	7,60
941	training	616	hexadecanoic acid	Ácido orgânico	57-10-3	5,277	7,23
942	training	617	11H-benzo[a]fluorene	HPA	238-84-6	4,315	5,46
943	training	618	11H-benzo[b]fluorene	HPA	243-17-4	4,505	5,31
944	training	619	ciprofloxacin	Outros compostos	85721-33-1	4,79	-0,57
945	test	308	morphine	Outros compostos	57-27-2	1,829	0,99
946	training	620	napropamide	Amida	15299-99-7	2,54	3,43
947	test	309	undecylbenzene	Benzeno e Alquil benzeno	6742-54-7	5,805	8,02
948	test	310	chrysene	HPA	218-01-9	4,494	5,71
949	test	311	benz[a]anthracene	HPA	56-55-3	4,592	5,72
950	test	312	naphthacene	HPA	92-24-0	4,51	5,71
951	test	313	triphenylene	HPA	217-59-4	4,364	5,77
952	training	621	2,2′-biquinoline	Heterociclo poliaromático	119-91-5	4,26	4,31
953	test	314	6-chrysenamine	HPA	2642-98-0	5,58	4,81
954	training	622	p-terphenyl	Derivados benzênicos	92-94-4	4,657	5,70
955	training	623	triphenylamine	Derivados benzênicos	603-34-9	4,5	5,03
956	training	624	triphenyl phosphate	Organofosforados	115-86-6	3,874	4,16
957	training	625	triphenylphosphine	Organofosforados	603-35-0	4,472	5,46
958	training	626	dicumyl peroxide	Outros compostos	80-43-3	4,369	5,43
959	training	627	dodecylbenzene	Benzeno e Alquil benzeno	123-01-3	6,083	8,38
960	training	628	linolenic acid	Ácido orgânico	463-40-1	4,891	6,62
961	test	315	linoleic acid	Ácido orgânico	60-33-3	5,212	7,06
962	training	629	oleic acid 2027-47-6	Ácido orgânico	112-80-1	5,533	7,68
963	training	630	octadecanoic acid	Ácido orgânico	57-11-4	5,854	8,02
964	test	316	fluridone	Heterociclo aromático	59756-60-4	3,01	4,11
965	training	631	triphenylmethanol	Derivados benzênicos	76-84-6	3,379	4,31
966	training	632	pencycuron	Outros compostos	66063-05-6	3,33	4,77
967	training	633	enrofloxacin	Outros compostos	93106-60-6	4,85	0,58
968	training	634	tridecylbenzene	Benzeno e Alquil benzeno	123-02-4	6,469	8,63
969	training	635	perylene	HPA	198-55-0	4,777	6,34
970	training	636	benzo[a]pyrene	HPA	50-32-8	4,75	6,39
971	test	317	13H-dibenzo[a,i]carbazole	Heterociclo poliaromático	239-64-5	6,1	6,10
973	training	637	7,12-dimethylbenz[a]anthracene	HPA	57-97-6	5,29	6,61
974	training	638	5,8,11,14-eicosatetraenoic acid	Ácido orgânico	506-32-1	5,174	6,80
975	training	639	eicosanoic acid	Ácido orgânico	506-30-9	6,431	8,53
976	training	640	1,2-dihydro-3-methylbenz[j]aceanthrylene	HPA	56-49-5	6,1	6,49
978	training	641	cis-permethrin	Outros compostos	61949-76-6	3,19	6,24
982	test	318	trans-permethrin	Outros compostos	61949-77-7	3,19	6,24
985	test	319	strychnine	Outros compostos	57-24-9	4,14	1,68
986	test	320	benzo[ghi]perylene	HPA	191-24-2	5,131	7,11
987	training	642	dibenz[a,h]anthracene	HPA	53-70-3	6,07	6,93
988	test	321	coronene	HPA	191-07-1	4,668	7,26
989	training	643	dioctyl phthalate	Éster	117-84-0	5,511	7,76

Nota: Adaptado de Shao et al. (2014).

Tabela S2 Critérios de classificação dos compostos químicos

Continua

Grupo	Estrutura Geral	Comentários
Ácido orgânico (G1)		R = grupo alquil com substituinte halogênio, grupo éster, OH, com SH, aminoácidos, anel aromático ou CN.
Alcano (G2)	-(CH2)n-	De cadeia aberta e fechada.
Alcano halogenado (G3)	 (X= F, Cl, Br, I)	De cadeia aberta e fechada.
Alceno halogenado (G4)	R-CH=CH-RX (X= F, Cl, Br, I)	De cadeia aberta e fechada.
Alcenos e alcinos (G5)	R-CH=CH-R R-C≡C-R	De cadeia aberta e fechada. R = grupo alquil com substituinte etoxi.
Álcool (G6)		R = grupo alquil, alceno ou alcino com halogênio, éter.
Amida (G7)		De cadeia aberta ou fechada (lactama). R = alquil com halogênio, com substituinte no nitrogênio, anel aromático.
Amina (G8)		R = grupo alquil com substituinte halogênio, hidroxila, com substituinte no nitrogênio.
Anilinas (G9)		R = grupo alquil com substituinte halogênio, hidroxila, com substituinte no nitrogênio.
Benzeno e Alquil benzeno (G10)		R = grupo alquil ou hidrogênio.
Benzeno halogenado (G11)	 (X= F, Cl, Br, I)	R = substituinte alquil.
Bifenil (G12)		R = grupo alquil, halogênio, amina, hidroxila (anéis com mais de um substituinte).
Compostos carbonílicos (G13)	 Aldeído cetona	De cadeia aberta ou fechada. R = grupo alquil, alceno ou alcino com substituinte halogênio.
Derivados benzênicos (G14)		Y = substituinte alquil, alceno, álcool, amina, éter, nitrila, aldeído, cetona, carbamatos (mais de um substituinte Y).
Éster (G15)		De cadeia aberta e fechada (lactonas). R = grupo alquil com ácido carboxílico, com anel aromático.

Grupo	Estrutura Geral	Comentários
Éter (G16)	-CH_2-O-CH_2-	Cíclico e de cadeia aberta.
Fenil ureia (G17)		R = grupo alquil.
Fenóis (G18)		R = grupo alquil com substituinte halogênio, nitro, hidroxila, metoxila e outros aneis ligados ao benzeno.
Heterociclo (G19)		Compostos cíclicos com ao menos um átomo diferente de C no anel (por exemplo: dioxano, trioxano, pirrolidina, morfolina).
Heterociclo aromático (G20)		Azois, piridina, tiazol, com substituintes alquil, halogênio, hidroxila.
Heterociclo poliaromático (G21)		Com substituintes alquil, halogênio, hidroxila.
HPA (Hidrocarboneto Policíclico Aromático) (G22)		Sem e com substituintes alquil, OH.
Nitrila (G23)	R-C≡N	R = grupo alquil com substituinte OH.
Nitroalcano (G24)	R-NO_2	R = grupo alquil.
Nitrobenzeno (G25)		R = grupo alquil com subsbstituinte halogênio.
Organofosforados (G26)		Qualquer composto que tenha um átomo de fósforo (fosfatos, fosfinas.
Organossulfurado (G27)		Qualquer composto que tenha ao menos um átomo de enxofre (tióis, tioesteres, tioeteres, dissulfetos, derivados sulfonas, derivados sulfito).
Triazinas (G28)		Pode ser a 1, 2,3; 1,2,4 ou 1,3,5 com substituintes diversos no anel.
Outros compostos (G29)		Triol, trinitrato, hidrazinas, oximas, siloxanos, benzotiazol, benzimidazol, benzodioxol, diazo, açúcar.

Observações sobre a Tabela S2:

1. Os 964 compostos foram organizados em 29 grupos de acordo com as semelhanças estruturais. Para aqueles compostos que apresentaram em suas estruturas mais de um grupo funcional, foi considerado um como o grupo principal, como, por exemplo, o caso dos ésteres, fenóis, nitrilas, álcoois, aminas, anilinas, nitrobenzeno e ácidos orgânicos.
2. Para os compostos classificados como derivados benzênicos foram consideradas estruturas que possuíam ao menos um anel aromático em sua estrutura, independentemente do número de substituintes.

3. Para os compostos classificados como organofosforados foi considerada a estrutura que apresentou ao menos um átomo de fósforo, independente do seu grupo funcional (fosfato, fosfina, etc.).
4. Para os compostos organossulfurados, foi considerada a presença de ao menos um átomo de enxofre, independente do grupo a que ele pertence (tiol, tio éster, sulfona, etc.).
5. Em alguns casos, quando o número de determinado composto foi muito pequeno, ele foi colocado num grupo que mais se assemelhava em termos de polaridade, como o caso dos alcenos e alcinos.
6. O grupo denominado “Outros compostos” foi constituído por compostos de estruturas muito diversificadas.

Tabela S3 Quantidade de compostos por classe, nos conjuntos de treinamento e no conjunto de teste

	A	H1	H2	Q1	Q2	Q3	Q4	E1	E2	E3	E4	E5	E6	E7	E8	TST
G1=Ácido Orgânico	48	24	24	13	9	11	15	6	4	4	8	7	5	7	7	19
G2=Alcano	16	7	9	4	3	3	6	2	2	2	2	2	1	1	4	10
G3=Alcano Halogenado	43	21	22	13	14	8	8	9	8	5	4	4	6	3	4	21
G4=Alceno Halogenado	9	6	3	3	2	3	1	1	1	3	---	2	1	---	1	4
G5=Alcenos e Alcinos	29	14	15	4	6	10	9	3	2	3	3	1	4	7	6	9
G6=Álcool	37	19	18	9	12	10	6	4	5	4	3	5	7	6	3	16
G7=Amida	13	6	7	2	4	4	3	1	2	1	1	1	2	3	2	3
G8=Amina	19	8	11	3	4	5	7	2	3	4	4	1	1	1	3	11
G9=Anilinas	28	13	15	8	8	5	7	5	3	2	4	3	5	3	3	16
G10=Benzeno e Alquil Benzeno	23	11	12	5	4	6	8	3	1	2	4	2	3	4	4	13
G11=Benzeno Halogenado	18	9	9	5	5	4	4	2	1	2	2	3	4	2	2	13
G12=Bifenil	27	13	14	8	7	5	7	5	3	2	3	3	4	3	4	14
G13=Compostos Carbonílicos	26	13	13	7	6	6	7	1	1	2	5	6	5	4	2	9
G14=Derivados Benzênicos	68	36	32	16	16	20	16	7	9	14	8	9	7	6	8	17
G15=Éster	31	16	15	4	10	12	5	1	6	8	2	3	4	4	3	14
G16=Éter	12	6	6	5	1	1	5	---	1	---	4	5	---	1	1	8
G17=Fenil Ureia	21	12	9	6	6	6	3	2	4	2	3	4	2	4	---	3
G18=Fenóis	36	20	16	8	8	12	8	4	4	6	2	4	4	6	6	30
G19=Heterociclo	6	4	2	2	---	2	2	1	---	1	2	1	---	1	---	4
G20=Heterociclo Aromático	19	6	13	2	7	4	6	2	4	3	3	---	3	1	3	16
G21=Heterociclo Poliaromático	6	3	3	1	2	2	1	1	2	1	---	---	---	1	1	8
G22=HPA	29	13	16	9	8	4	8	5	4	1	5	4	4	3	3	20
G23=Nitrila	10	7	3	4	3	3	---	2	1	2	---	2	2	1	---	6
G24=Nitroalcano	5	2	3	---	---	2	3	---	---	1	2	---	---	1	1	1
G25=Nitrobenzeno	10	4	6	4	2	---	4	3	2	---	4	1	---	---	---	12
G26=Organo Fosforado	17	7	10	7	5	---	5	4	3	---	1	3	2	---	4	2
G27=Organossulfurado	9	7	2	3	1	4	1	2	---	2	---	1	1	2	1	9
G28=Triazinas	4	3	1	1	1	2	---	1	---	---	---	---	1	2	---	3
G29=Outros Compostos	20	9	11	4	6	5	5	2	3	3	1	2	3	2	4	10
Total	**639**	**319**	**320**	**160**	**160**	**159**	**160**	**81**	**79**	**80**	**80**	**79**	**81**	**79**	**80**	**321**

Tabela S4 Valores mínimos/máximos de log K_{oc} experimental por classe de compostos, nos conjuntos de treinamento e no conjunto de teste

Continua

	A	H1	H2	Q1	Q2	Q3	Q4	E1	E2	E3	E4	E5	E6	E7	E8	TST
G1=Ácido Orgânico	-0,282	-0,282	0,441	-0,282	1,497	1,250	1,600	1,121	1,883	1,250	1,600	-0,282	1,497	1,502	1,083	-0,630
	6,431	6,431	5,854	5,533	5,854	6,431	5,277	5,533	5,854	6,431	4,891	3,602	3,860	2,574	5,277	5,212
G2=Alcano	1,970	1,970	2,628	1,970	2,628	2,362	2,949	2,661	3,488	2,362	2,949	1,970	2,628(*)	3,069(*)	3,009	2,313
	4,695	3,553	4,695	3,254	4,113	3,553	4,695	3,254	4,113	3,553	3,825	3,248			4,695	5,294
G3=Alcano Halogenado	0,790	0,790	1,650	0,790	1,650	1,654	1,800	0,790	1,650	1,654	2,057	1,486	2,019	2,010	1,800	1,040
	3,749	3,635	3,749	3,553	3,749	3,635	3,210	3,553	3,749	3,635	3,118	3,300	3,444	3,096	3,210	3,410
G4=Alceno Halogenado	2,052	2,128	2,052	2,128	2,052	2,150	2,310(*)	2,481(*)	2,052(*)	2,150	---	2,128	2,427(*)	---	2,310(*)	2,351
	4,110	4,110	2,427	2,536	2,427	4,110				4,110		2,536				4,200
G5=Alcenos e Alcinos	1,943	1,943	2,166	1,943	2,171	2,460	2,166	1,943	2,634	2,655	2,693	2,645(*)	2,171	2,460	2,166	1,578
	4,179	4,119	4,179	2,933	3,227	4,119	4,179	2,933	3,227	4,005	3,825		2,900	4,119	4,179	3,852
G6=Álcool	0,458	0,458	0,637	0,572	0,637	0,458	1,083	0,958	0,925	1,513	1,404	0,572	0,637	0,458	1,083	0,180
	4,657	4,657	3,564	4,657	3,564	1,393	2,378	2,802	3,564	3,863	2,073	4,657	3,047	1,393	2,378	2,111
G7=Amida	0,556	0,556	0,806	0,692	0,806	0,556	0,958	0,692(*)	0,849	0,556(*)	2,540(*)	0,828(*)	0,806	0,915	0,958	1,442
	2,540	1,170	2,540	0,828	1,274	1,170	2,540		1,263				1,274	1,170	1,943	2,802
G8=Amina	0,267	0,866	0,267	0,866	0,267	1,464	0,833	0,866	0,267	1,464	0,833	1,518(*)	1,595(*)	2,188(*)	1,638	0,599
	2,955	2,955	2,917	1,774	2,285	2,955	2,917	1,774	2,285	2,955	2,917				2,775	2,720
G9=Anilinas	0,670	0,670	0,670	0,670	0,670	0,670	0,670	0,670	1,214	1,740	0,670	1,459	0,670	0,670	1,197	0,670
	4,490	4,490	3,130	4,490	2,383	3,130	3,130	4,490	2,383	1,960	1,900	2,360	2,360	3,130	3,130	3,178
G10=Benzeno e Alquil Benzeno	1,970	3,239	1,970	3,237	3,297	3,352	1,970	3,542	3,781(*)	3,863	3,368	3,237	3,297	3,352	1,970	1,870
	6,083	6,083	4,380	3,809	3,852	6,083	4,380	3,809		4,043	4,380	3,607	3,852	6,083	3,825	5,805
G11=Benzeno Halogenado	2,220	2,220	2,536	2,220	2,600	2,780	2,536	2,600	2,764(*)	3,417	2,536	2,220	2,600	2,780	2,612	2,579
	4,490	4,490	3,520	4,490	3,237	3,520	3,520	2,628		3,520	3,520	4,490	3,237	2,850	2,781	4,113
G12=Bifenil	3,830	3,830	3,830	3,830	3,830	3,830	3,830	4,146	3,830	3,830	3,836	3,830	3,885	4,700	3,830	2,690
	5,816	5,816	5,640	5,816	5,640	5,640	5,210	5,816	5,210	5,210	5,210	5,370	5,640	5,640	5,000	5,870
G13=Compostos Carbonílicos	0,630	1,535	0,630	1,595	0,630	1,535	1,372	2,345(*)	1,682(*)	1,693	1,372	1,595	0,630	1,535	2,111	0,630
	3,607	3,607	3,428	2,965	2,400	3,607	3,428			1,698	2,666	2,965	2,400	3,607	3,428	3,602
G14=Derivados Benzênicos	1,570	1,709	1,570	1,997	1,570	1,709	1,950	2,030	1,570	1,790	1,950	1,997	1,660	1,709	2,171	1,453
	6,469	4,804	6,469	3,667	6,469	4,804	5,245	2,509	4,657	4,500	5,245	3,667	6,469	4,804	3,994	3,977
G15=Éster	1,029	1,029	1,263	1,029	1,829	1,513	1,263	2,035(*)	1,829	1,513	1,263	1,029	2,079	2,035	1,899	0,485
	5,511	5,511	3,466	2,944	3,466	5,511	2,574		3,466	3,537	2,574	2,944	2,802	5,511	2,367	3,776
G16=Éter	1,181	1,181	1,393	1,214	1,682(*)	1,181(*)	1,393	---	1,682(*)	---	1,393	1,214	---	1,181(*)	3,123(*)	1,263
	3,123	2,035	3,123	2,035			3,123				2,481	2,035				2,481

	A	H1	H2	Q1	Q2	Q3	Q4	E1	E2	E3	E4	E5	E6	E7	E8	TST
G17=Fenil Ureia	1,310 2,820	1,310 2,820	1,430 2,580	1,310 2,820	1,430 2, 430	1,350 1,960	2,120 2,580	2,350 2,820	1,430 2, 070	1,520 1,600	2,120 2,580	1,310 2,560	1,700 2, 430	1,350 1,960	---	2,010 2, 430
G18=Fenóis	1,550 3,760	1,690 3,760	1,550 3,760	1,690 3,760	1,550 3,760	1,710 3,760	1,960 3,760	2,610 3,118	1,550 2,550	1,710 2,737	2,720 2,737	1,690 3,760	2,610 3,760	2,410 3,760	1,960 3,760	1,020 4,151
G19=Heterociclo	0,398 1,834	0,398 1,834	1,149 1,741	0,398 0,741	---	1,627 1,834	1,149 1,741	0,741(*)	---	1,834(*)	1,149 1,741	0,398(*)	---	1,627(*)	---	1,143 2,383
G20=Heterociclo Aromático	1,159 3,129	1,448 2,645	1,159 3,129	2,030 2,291	1,333 2,786	1,448 2,645	1,159 3,129	2,030 2,291	1,333 2,383	1,448 2,645	1,785 3,129	---	1,442 2,786	1,616(*)	1,159 1,627	1,529 4,130
G21=Heterociclo Poliaromático	2,476 4,260	3,227 4,260	2,476 3,074	4,260(*)	2,476 3,074	3,227 3,401	2,481(*)	4,260(*)	2,476 3,074	3,227(*)	---	---	---	3,401(*)	2,481(*)	2,373 6,100
G22=HPA	2,220 6,100	3,000 5,290	2,220 6,100	3,000 5,290	2,922 6,070	3,482 4,777	2,220 6,100	3,722 5,290	2,922 4,472	3,722(*)	3,580 6,100	3,000 4,315	3,325 6,070	3,482 4,777	2,220 4,750	2,220 5,580
G23=Nitrila	0,985 2,873	0,985 2,247	1,627 2,873	0,985 2,247	1,627 2,873	1,203 1,747	---	1,703 2,247	2,873(*)	1,415 1,747	---	0,985 1,501	1,627 1,888	1,203(*)	---	0,838 1,595
G24=Nitroalcano	1,197 2,470	1,850 2,177	1,197 2,470	---	---	1,850 2,177	1,197 2,470	---	---	2,177(*)	1,197 2,470	---	---	1,850(*)	1,883(*)	1,475(*)
G25=Nitrobenzeno	2,010 2,715	2,296 2,596	2,010 2,715	2,296 2,596	2,010 2,420	---	2,019 2,715	2,296 2,596	2,010 2,420	---	2,019 2,715	2,420(*)	---	---	---	2,171 2,802
G26=Organo Fosforado	1,023 5,070	1,640 5,070	1,023 3,874	1,640 5,070	1,023 3,510	---	2,470 3,874	1,812 5,070	2,360 3,510	---	2,500(*)	1,640 4,472	1,023 1,529	---	2,470 3,874	2,520 2,820
G27=Organossulfurado	0,610 3,010	0,610 3,010	1,589 2,008	0,643 2,340	1,589(*)	0,610 3,010	2,008(*)	1,300 2,340	---	1,997 3,010	---	0,643(*)	1,589(*)	0,610 2,215	2,008(*)	0,900 3,270
G28=Triazinas	2,130 3,390	2,130 3,390	2,430(*)	3,390(*)	2,430(*)	2,130 2,800	---	3,390(*)	---	---	---	---	2,430(*)	2,130 2,800	---	2,190 2,740
G29=Outros Compostos	-0,386 5,150	-0,386 5,150	2,040 5,150	-0,386 4,151	2,106 5,150	1,834 5,150	2,040 3,934	-0,386 4,151	2,106 3,178	1,986 3,379	3,934(*)	2,260 3,190	2,976 5,150	1,834 5,150	2,040 3,330	1,829 5,150

(*) 1 composto.