**UNIVERSIDADE FEDERAL DE GOIÁS**
**FACULDADE DE FARMÁCIA**

**GABRIELLE SANTOS RAMOS**

# Modelos de aprendizado profundo para avaliação de toxicidade aguda de compostos químicos em aves

Goiânia
2022

UNIVERSIDADE FEDERAL DE GOIÁS
FACULDADE DE FARMÁCIA

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)**

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Gabrielle Santos Ramos

Título do trabalho: Modelos de aprendizado profundo para avaliação de toxicidade aguda de compostos químicos em aves

**2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento** [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

**Casos de embargo:**

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

**Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.**

| | |
|---|---|
| | Documento assinado eletronicamente por **Bruno Junior Neves**, **Professor do Magistério Superior**, em 24/08/2022, às 15:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020. |

| | |
|---|---|
| | Documento assinado eletronicamente por **GABRIELLE SANTOS RAMOS**, **Discente**, em 04/09/2022, às 21:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020. |

| | |
|---|---|
| | A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3083255** e o código CRC **BF32B0A9**. |

**GABRIELE SANTOS RAMOS**

# Modelos de aprendizado profundo para avaliação de toxicidade aguda de compostos químicos em aves

Trabalho de Conclusão de Curso apresentado ao curso de Farmácia da Universidade Federal de Goiás, como requisito parcial para a obtenção do título de Bacharel em Farmácia.

Orientador(a): Prof. Dr. Bruno Junior Neves

Goiânia

2022

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Ramos, Gabrielle Santos Ramos
Modelos de aprendizado profundo para avaliação de toxicidade aguda de compostos químicos em aves [manuscrito] / Gabrielle Santos Ramos Ramos. - 2022.
33 f.

Orientador: Prof. Dr. Bruno Junior Neves; co-orientador Meryck Filipe Brito.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Faculdade Farmácia (FF), Farmácia, Goiânia, 2022.
Bibliografia.
Inclui gráfico, tabelas, lista de figuras, lista de tabelas.

1. ecotoxicologia . 2. multitask . 3. QSAR. I. Neves, Bruno Junior, orient. II. Título.

CDU 615.1

FACULDADE DE FARMÁCIA

# ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e quatro dias do mês de agosto do ano de 2022 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Modelos de aprendizado profundo para avaliação de toxicidade aguda de compostos químicos em aves”, de autoria de Gabrielle Santos Ramos, do curso de Farmácia, da Faculdade de Farmácia da UFG. Os trabalhos foram instalados pelo Prof. Dr. Bruno Junior Neves – orientador - Faculdade de Farmácia/UFG, com a participação dos demais membros da Banca Examinadora: Me. Vinicius Alexandre Fiaia Costa (Faculdade de Farmácia/UFG) e Dra. Eufrásia de Sousa Pereira (Faculdade de Farmácia/UFG). Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9,5 pontos, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

| | |
|---|---|
| sei! assinatura eletrônica | Documento assinado eletronicamente por **Bruno Junior Neves**, **Professor do Magistério Superior**, em 24/08/2022, às 15:28, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020. |
| sei! assinatura eletrônica | Documento assinado eletronicamente por **Eufrasia de Sousa Pereira**, **Usuário Externo**, em 08/09/2022, às 15:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020. |
| sei! assinatura eletrônica | Documento assinado eletronicamente por **VINÍCIUS ALEXANDRE FIAIA COSTA**, **Usuário Externo**, em 12/09/2022, às 09:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020. |
| | A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3083246** e o código CRC **AF654D1E**. |

## RESUMO

A modernização da agricultura tem proporcionado crescimento econômico em decorrência do aumento na produtividade. Todavia este setor tem como prática o uso intenso de agrotóxicos, que apresenta risco potencial ao meio ambiente e organismos prestadores de serviços ecológicos como as aves. Considerando que as aves são organismos dispersores de sementes, estudos de toxicidade aguda *in vivo* têm sido amplamente utilizados como critério regulatório para o registro de novos agrotóxicos. Entretanto, estes estudos geralmente demandam muito tempo, apresentam alto custo e envolvem questões éticas. Diante disso, este trabalho teve como objetivo desenvolver modelos de Relação Quantitativa entre Estrutura e Atividade - *Quantitative Structure Activity Relationship* (QSAR), baseados em aprendizado de máquina, para predizer a toxicidade aguda de compostos químicos em diversas espécies de aves. Inicialmente foi realizada a compilação, integração e preparação dos maiores conjuntos de dados de compostos com dados de propriedades toxicológicas experimentais para as seguintes espécies de aves: *A. platyrhynchos, C.virginianus, C. japônica* e *P. colchicus*. Em seguida, uma análise de espaço químico demonstrou que os conjuntos de dados preparados compartilham informação química entre si, sendo que a correlação dos dados toxicológicos entre as espécies demonstrou-se moderada (sendo 'r' em torno de 0,68). Ao final deste processo, modelos de QSAR para tarefas de regressão foram gerados utilizando métodos de *Deep Learning*. Entre eles, um modelo *multitask* baseado em redes neurais do tipo *Feed-Forward Neural Networks* (FFNN) ,capaz de predizer a toxicidade aguda ($pDL_{50}$) de várias espécies de ave de forma simultânea, foi o mais preditivo, obteve-se valores de *r* entre 0,59 – 0,80 para o conjunto teste. Os resultados demonstram que o modelo multitarefa foi capaz de promover transferência indutiva de aprendizado entre as *tasks*, ou seja, entre os bioensaios de cada espécie. Portanto,o modelo gerado representa um novo método alternativo ao uso de animais para avaliação de toxicidade aguda em aves.



palavras-chave: ecotoxicologia, *multitask* , QSAR

**ABSTRACT**

The modernization of agriculture has provided economic growth as a result of increased productivity. However, this sector has the intense use of pesticides as a practice, which presents a potential risk to the environment and organisms that provide ecological services such as avian species. Considering that birds are seed-dispersing organisms, in vivo acute toxicity studies have been widely used as regulatory criteria for the registration of new pesticides. However, these studies are usually time consuming, expensive and involve ethical issues. Therefore, this work aimed to develop Quantitative Structure Activity Relationship (QSAR) models, based on machine learning, to predict the acute toxicity of chemical compounds in several bird species. Initially, the compilation, integration and preparation of the largest datasets of compounds with data on experimental toxicological properties were performed for the following avian species: *A. platyrhynchos, C.virginianus, C. japonica* and *P. colchicus*. Then, a chemical space analysis showed that the prepared datasets share chemical information with each other, and the correlation of toxicological data between species proved to be moderate (with 'r' around 0.68). At the end of this process, QSAR models for regression tasks were generated using Deep Learning methods. Among them, a multitask model based on Feed-Forward Neural Networks (FFNN), capable of predicting the acute toxicity (pDL50) of several bird species simultaneously, was the most predictive, obtaining r values between 0.59 – 0.80 for the test set. The results demonstrate that the multitasking model was able to promote inductive transfer of learning between tasks, that is, between bioassays of each species. Therefore, the generated model represents a new alternative method to the use of animals for the evaluation of acute avian toxicity.

keywords: ecotoxicology, multitask, QSAR

**SUMÁRIO**

## LISTA DE FIGURAS



## LISTA DE TABELAS

# 1. INTRODUÇÃO

A agricultura constitui uma atividade importante desde os primórdios na história da humanidade, e tem se desenvolvido significativamente no decorrer dos anos. No Brasil, esta prática se destaca como uma das principais bases econômicas do país. Apesar de sua importância na economia brasileira e para a manutenção de seguridade alimentar global, o uso extensivo de agrotóxicos representa uma ameaça para organismos prestadores de serviços ecológicos, tais como abelhas, peixes e aves.[1]

Dentre os organismos que atuam como bioindicadores, destacam-se as aves. Estes animais são responsáveis pelo controle de insetos e dispersão de sementes, proporcionando desta forma, a manutenção da biodiversidade da flora [2]. Apesar disso, esses animais acabam sendo expostos a concentrações letais e subletais de agrotóxicos, que podem promover malformações, redução da fecundidade, distúrbios hormonais, alterações no senso de direção, prejudicando estas espécies não-alvo.[3]

Diante deste contexto, agências governamentais e órgãos consultivos científicos tem buscado desenvolver e padronizar métodos e estratégias para avaliar os efeitos tóxicos de agrotóxicos e produtos químicos industriais em aves. Atualmente, a Organização para Cooperação e Desenvolvimento Econômico (OECD) [4] recomenda a avaliação experimental em espécies modelo de aves, com o intuito de determinar a dose letal que induz a morte de 50% de uma população de animais ($LD_{50}$) e/ou a concentração letal que induz a morte de 50% de uma população de animais ($LC_{50}$). Entretanto, a avaliação completa da toxicidade para uma grande quantidade de substâncias químicas (em várias doses e concentrações) por meio destes ensaios experimentais é demorada, dispendiosa e representa um problema ético. Portanto, o desenvolvimento de novos métodos alternativos ao uso de animais para avaliar a toxicidade aguda de produtos químicos em aves é urgentemente necessário.[5]

Diante desta emergência, abordagens de aprendizado de máquina, que constituem um dos componentes mais importantes do campo da inteligência artificial, têm sido utilizadas para encontrar relações quantitativas entre estrutura química e toxicidade a partir de conjuntos de dados de compostos com toxicidade experimental determinada. Metodologicamente, a modelagem dos parâmetros toxicológicos pode ser representada como um processo de três partes. Inicialmente, as estruturas químicas de compostos são convertidas matematicamente em valores numéricos ou sistemas matriciais por meio de métodos computacionais

conhecidos como descritores moleculares (variáveis independentes), o que torna a informação química um valor compreensível para modelagem computacional. Em seguida, métodos de aprendizado de máquina *(Random Forest, Deep Neural Networks, Support Vector Machine, etc*) são utilizados para estabelecer relações quantitativas entre os descritores moleculares e propriedade toxicológica estudada (variável dependente). Esta etapa envolve a utilização de uma função estabelece pesos aos descritores moleculares, conforme a seguinte equação:

$$P(t) = k'\left(D_1 D_2 \vdots D_n\right)$$

Sendo que Pt é definido como propriedade toxicológica das moléculas, $D_1$, $D_2$,…, Dn representam os descritores moleculares, e k' é o peso estabelecido pelo algoritmo selecionado. Ao final deste processo, a preditividade do modelo obtido é calculada utilizando métricas estatísticas, as quais irão avaliar a capacidade do modelo em predizer corretamente a toxicidade de compostos avaliados experimentalmente.[6,7] Uma vez validado, o modelo gerado representa uma ferramenta inestimável para a otimizar fluxos regulatórios para o registro e reavaliação de agrotóxicos com perfis ecotoxicológicos mais adequados.

Diante do exposto, o objetivo geral deste trabalho consiste em desenvolver modelos de regressão robustos e preditivos de QSAR baseados em aprendizado de máquina para avaliação de toxicidade aguda de agrotóxicos em aves. Os objetivos específicos são: (i) compilar, integrar e preparar os conjuntos de dados obtidos de compostos com propriedades toxicológicas experimentais para organismos não alvo (aves); (ii) analisar o espaço químico e a diversidade estrutural dos conjuntos de dados preparados; (iii) construir e validar modelos contínuos por meio de diversas abordagens e algoritmos de aprendizado de máquina.

## 2. MATERIAIS E MÉTODOS

### 2.1. Preparo dos conjuntos de dados

Inicialmente, todos os compostos químicos com *endpoints* toxicológicos $LD_{50}$ (dose letal para 50% da população) para espécies de aves foram coletados a partir das seguintes bases de dados de domínio público: ECOTOX,[8] EFSA[9] e revisão da literatura. Quatro espécies de aves foram selecionadas para coleta de dados por apresentarem quantidades significativas de compostos com dados experimentais de $LD_{50}$: *Anas platyrhynchos*, *Colinus virginianus*, *Coturnix japônica* e *Phasianus colchicus.* Em seguida, apenas os compostos com

dados de pureza acima de 80% e ensaios com tempo de exposição de 7-8 e 14-15 dias foram mantidos nos conjuntos de dados. Em paralelo, as unidades de medida de toxicidade aguda foram convertidas de da escala de ppm (partes por milhão) para mM (milimolar).

O preparo dos dados representa um processo fundamental para garantir a eficiência do modelo preditivo, tendo em vista sua capacidade de evitar a propagação de erros que porventura possam existir nos dados. [10] Logo, em seguida, as estruturas químicas e dados toxicológicos correspondentes foram padronizadas utilizando o pacote RDKit disponível em Phyton v.3.7. [12] através das seguintes etapas: adição de hidrogênios explícitos às estruturas químicas, remoção de sais, misturas, polímeros e compostos organometálicos e normalização de quimiotipos específicos, como anéis aromáticos e grupos nitro [11]. Logo após a padronização, realizou-se a remoção de duplicatas (*i.e.*, registros químicos idênticos repetidos no conjunto de dados) e *outliers* (dados discordantes, que diferem drasticamente dos demais) através do programa GraphPad Prism.[12] Conforme protocolo de Fourches, utilizando scripts compilados em python, considerou-se a discordância dos valores biológicos em 3 vezes o desvio padrão e o cálculo de Z-score ( que corresponde à medida de quantos desvios padrão um valor de amostra está acima ou abaixo da média aritmética.). Os detalhes do número de compostos com dados de toxicidade aguda experimental para cada espécie de ave estão descritos na **Tabela 1**.

**Tabela 1:** DADOS PRÉ E PÓS-CURAGEM. Número de compostos obtidos para cada espécie de ave antes e depois do preparo dos dados químicos e biológicos.

| Espécie | Dados brutos | Dados preparados |
|---|---|---|
| *Anas platyrhynchos* | 865 | 568 |
| *Colinus virginianus* | 931 | 546 |
| *Coturnix japônica* | 158 | 139 |
| *Phasianus colchicus* | 174 | 126 |

Entre as espécies com mais dados de bioensaios destacam-se a *Anas platyrhynchos* (pato real) com 568 compostos e *Colinus virginianus* (perdiz-da-virgínia) com 546 compostos. Já as demais espécies apresentam um número mais escasso de dados ,sendo *Coturnix japônica* com 139 compostos e *Phasianus colchicus* com 126 compostos.

## 2.2. Análise de relevância e compartilhamento de informações químicas

Ao final do processo de preparo dos dados, uma matriz de correlação dos dados de $LD_{50}$ entre as espécies de aves foi construída utilizando o coeficiente de correlação de Pearson. Subsequentemente, utilizou-se o algoritmo de *K-means* com objetivo de analisar grupos de estruturas químicas semelhantes *(scaffolds*). Este algoritmo calcula a média entre os pontos do conjunto de dados, com base na distância euclidiana entre eles, de modo que o centroide corresponde a média aritmética de todos os pontos do cluster, o que permite a classificação conforme o número de clusters que melhor represente os dados. São realizadas iterações até o momento em que a mudança de posição dos centroides é encerrada. Foi utilizado o UMAP (*Uniform Manifold Aproximation and Projection*) [20] como uma técnica de redução de dimensionalidade para aprendizado de máquina. Para representar visualmente os dados foi utilizada a ferramenta *t-Distributed Stochastic Neighbour Embedding* (t-SNE). [14]

## *2.3.* Descritores moleculares

Os *fingerprints* são descritores moleculares representados através de sequências binárias (*bits*) que codificam a presença ou ausência de subestruturas específicas em estruturas químicas. [15] Neste trabalho foram utilizados *fingerprints* circulares do tipo Morgan ECFP4 utilizando o programa de código aberto RDKit, executado em Python v.3.7. [16]

## 2.4. Construção de uma matriz multitarefa

Inicialmente, foi gerada uma matriz esparsa contendo todas as informações químicas (SMILES- *Simplified molecular-input line-entry system*- que constituem um tipo de notação linear que descreve compostos químicos) e toxicológicas (valores de $LD_{50}$) para as espécies de aves. Em seguida, todos os dados de $LD_{50}$ foram convertidos para a escala logarítmica negativa ($pLD_{50}$). A matriz esparsa foi então dividida em três partes, ou seja, conjuntos treinamento, validação e teste na proporção de 8:1:1, respectivamente. Duas abordagens de divisão foram utilizadas: *random split* e *scaffold split*. Em *random split*, a divisão dos dados é realizada de forma randômica, ao passo que em *scaffold split*, a divisão assegura que compostos com arcabouços moleculares idênticos não sejam mantidos entre os conjuntos treinamento, teste e validação. [17]

### 2.5.Preparo e construção dos modelos *Multitask* e *single tasks*

Ao final do processo de construção da matriz, modelos multitarefa e para tarefas únicas foram gerados utilizando redes neurais do tipo *Feed-Forward Neural Networks* (FFNN) implementadas no pacote TensorFlow v.2.8. [18] A rede foi otimizada variando-se os seguintes hiperparâmetros: função de ativação ('ReLU', 'SELU' e 'LeakyReLU'), número de camadas ocultas (1–7), número de neurônios em cada camada (5-50) e função de otimização (SGD, ADAM, NADAM e RMSprop).[6] Para evitar o sobre ajuste (*overfitting*) dos modelos, uma parada antecipada (*early stopping*) foi aplicada durante o treinamento do modelo, com intervalo de variação do learning rate entre 0,1-0,0000001 e até 1000 épocas.

### 2.6. Validação estatística dos modelos

A capacidade preditiva dos modelos contínuos foi baseada nas seguintes métricas estatísticas: coeficiente de *Pearson* (r), que estabelece correlação positiva ou negativa, podendo variar entre 1 e -1 ; o erro absoluto médio (MAE), que corresponde ao módulo da diferença entre o valor experimental e a média; o erro quadrático médio (MSE), que eleva o erro absoluto ao quadrado, sendo, portanto, uma métrica mais sensível ao erro; e a raiz quadrada do erro médio (RMSE), que corresponde ao desvio padrão da diferença ao quadrado entre o valor predito e o valor experimental observado.[19]

## 3. RESULTADOS E DISCUSSÃO

### 3.1. Conjunto de dados

No presente trabalho, modelos de aprendizado de máquina foram gerados utilizando métodos de aprendizado de máquina para predição de toxicidade aguda em 4 espécies de aves. Inicialmente, uma análise exploratória inicial dos valores $LD_{50}$ reportados para *Anas platyrhynchos*, *Colinus virginianus*, *Coturnix japônica* e *Phasianus colchicus* demonstrou que embora estes bioensaios toxicológicos sejam relacionados, apresentam correlação moderada (r ~0,67) entre si. Em outras palavras, a diversidade genética entre as quatro espécies implica em perfis de resposta toxicológicas variáveis, com certo grau de correlação. Os detalhes da matriz de correlação de Pearson entre as quatro espécies de aves estão descritos na ***Figura 1*** a seguir.

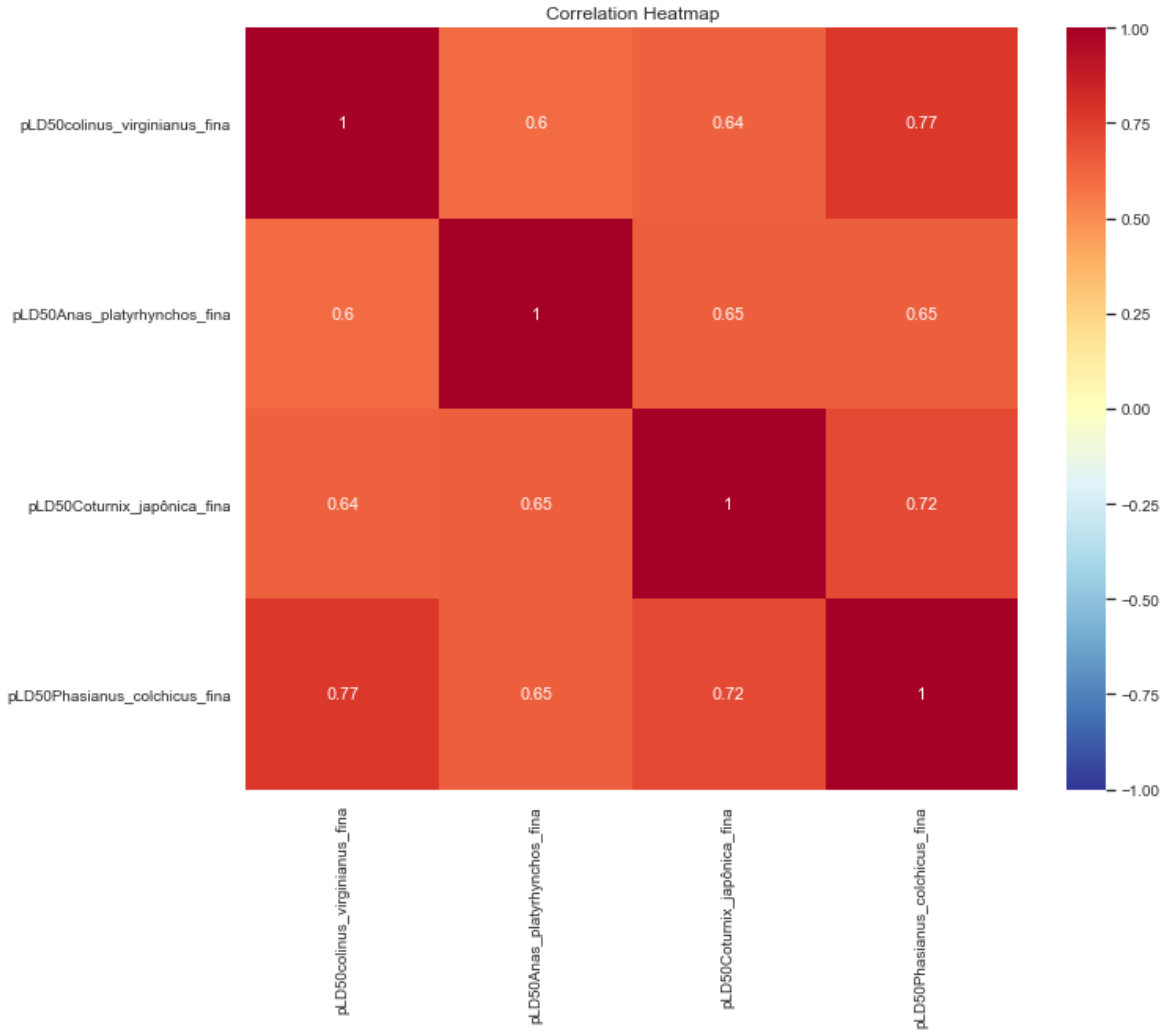



**Figura 1.** MATRIZ DE CORRELAÇÃO ENTRE $pLD_{50}$ DAS ESPÉCIES DE AVES.

A matriz entre os valores de $pLD_{50}$ para as 4 tasks apresentou correlação considerável entre os dados, com valores de correlação de pearson ≥ 0,6. (Sendo que quanto mais próximo de 1, maior é a correlação entre os dados.)

A análise do espaço químico foi realizada avaliando-se a similaridade estrutural entre as moléculas do conjunto de dados, cuja importância se deve à relação estabelecida entre a estrutura química e sua respectiva atividade biológica. Utilizou-se uma função de similaridade, coeficiente de Tanimoto, e por meio dos descritores ECFP4, validou-se o número de *clusters* com métrica de *Silhouette Coefficient* que calculou a semelhança entre os compostos químicos, agrupando em *clusters* , que mostram a distribuição dos dados no espaço. A diversidade estrutural do conjunto de dados pode ser visualizada através das figuras geradas pelo UMAP.

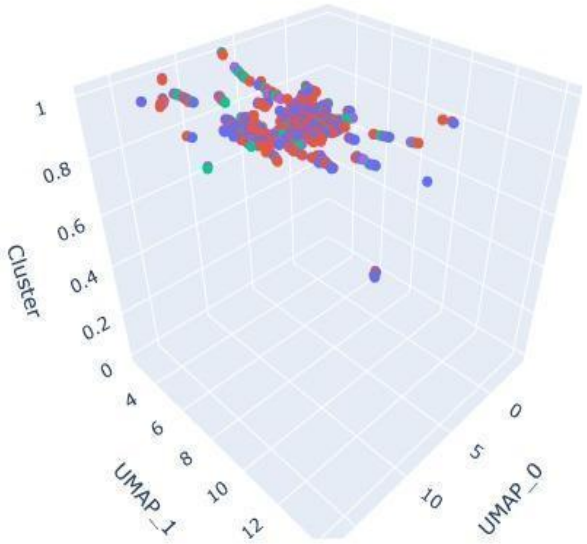



**Figura 2***:* *CLUSTERING* ENTRE *TASKS* -PELO ALGORITMO UMPA . Agrupamento de dados químicos compartilhados entre as diferentes *tasks.* Em azul - pLD50 *C.virginianus*; vermelho - pLD50 *A. platyrhynchos* ; verde- pLD50 *C. japônica;* roxo - pLD50 *P. colchicus*. (Autoria própria, 2022).

É possível visualizar os dados químicos dentro dos clusters, que foram validados utilizando-se um valor de silhouette de 0.7. A figura acima permite a visualização do espaço químico utilizando-se o algoritmo de *k-means*.

A figura a seguir permite a visualização sob diferentes ângulos da distribuição dos dados químicos dentro dos clusters.

Então, foi realizada análise química e biológica dos dados, através de busca na literatura para verificar a consistência dos dados experimentais, e identificou-se valores discordantes interlaboratoriais, caracterizados por compostos químicos cujas pequenas alterações nas suas estruturas moleculares demonstraram diferença brusca de atividade biológica entre si. Logo, estes dados discordantes foram excluídos do conjunto de dados pré-modelagem.

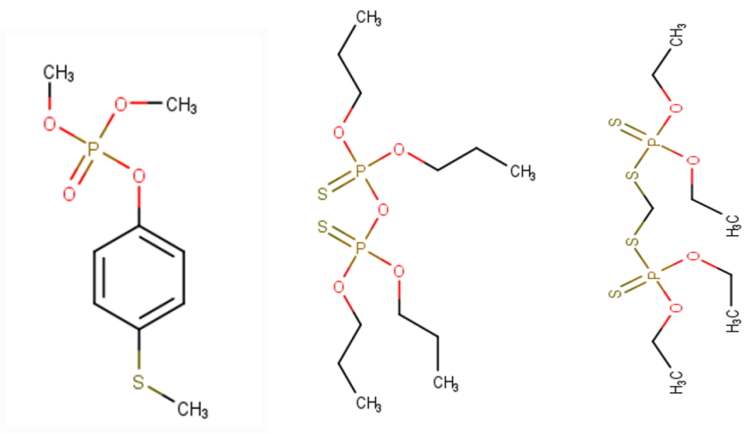


**Figura 3 :** ESTRUTURAS QUÍMICAS DISCORDANTES *(Autoria própria, 2022).*

A figura acima demonstra exemplos de estruturas químicas que apresentaram valores discrepantes de atividade. Observa-se que todas apresentam em comum o grupamento fosfato. Sendo este comum em pesticidas organofosforados, de caráter tóxico. É importante ressaltar que não foram removidas todas as estruturas químicas que apresentaram grupo fosfato em sua estrutura, o que logicamente comprometeria a capacidade preditiva do modelo, cujo objetivo consiste em prever toxicidade.

### *3.2.* Análise de *scaffolds*

A figura a seguir mostra os 10 *scaffolds* (“arcabouços moleculares”) mais presentes no conjunto de dados associados à respectiva *task* na qual se mostraram mais frequentes.

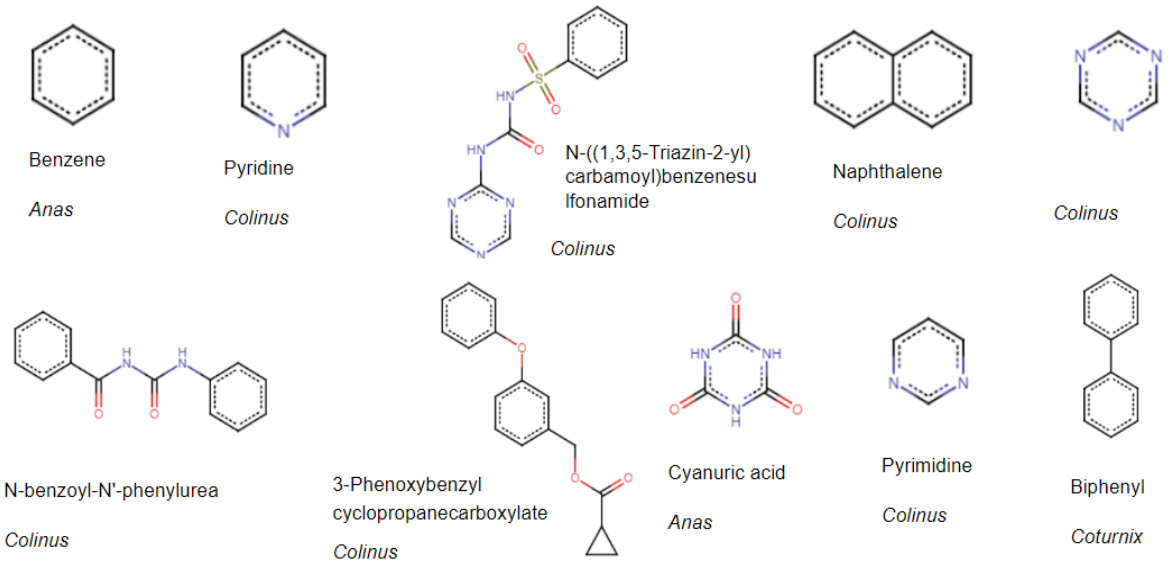


**Figura 4:** *SCAFFOLDS* - ARCABOUÇOS MOLECULARES MAIS FREQUENTES (Autoria própria, 2022).

Observa-se a presença significativa de anéis aromáticos e grupos nitrogenados nos arcabouços moleculares mais recorrentes na *dataset*. Logo, esses *scaffolds* são os mais representativos para o conjunto de dados.

### 3.3. *Deep learning*

A matriz esparsa é um meio de organizar os dados para adequá-los de modo a serem inseridos na rede de aprendizado. A seguinte figura demonstra a esparsidade dos dados entre as *tasks*:

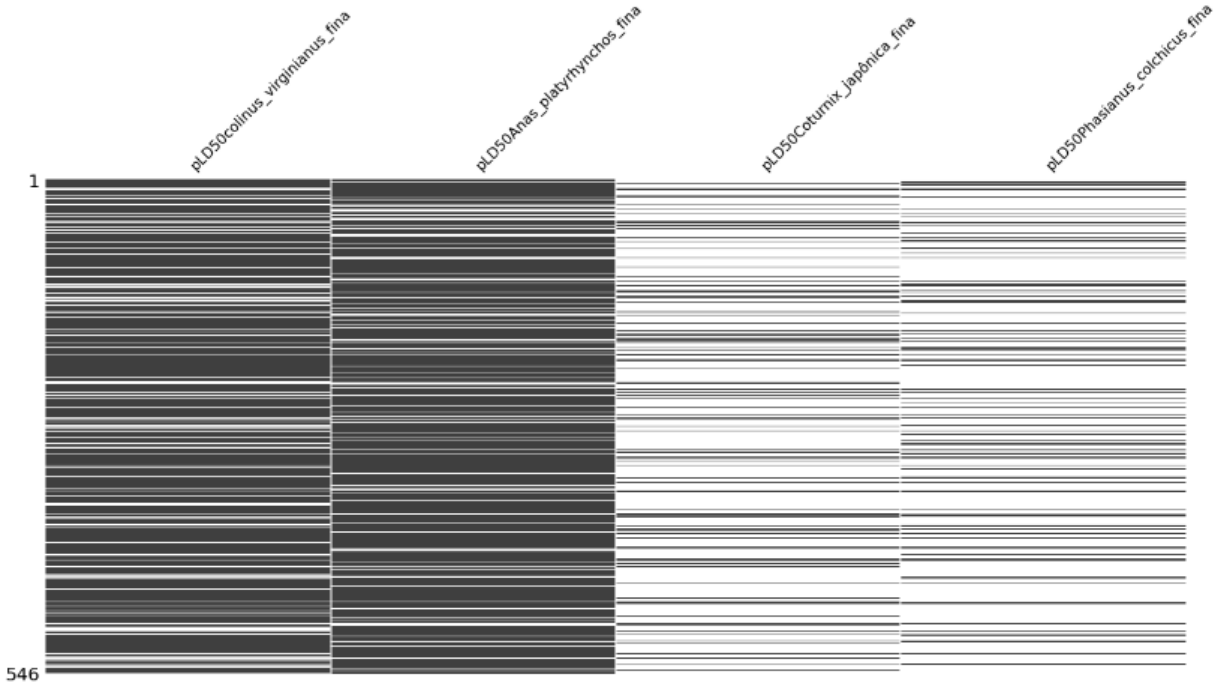



**Figura 5:** ESPARSIDADE DOS DADOS (Autoria própria, 2022).

A figura permite a visualização do compartilhamento de dados químicos entre as espécies de aves. Nota-se que os espaços vazios são maiores justamente entre as espécies (*tasks*) que apresentam menor número de dados. Visualmente as *tasks* 3 e 4 apresentam maior esparsidade. Já as *tasks* 1 e 2 apresentam maior densidade, ou seja, compartilham maior número de dados.

.

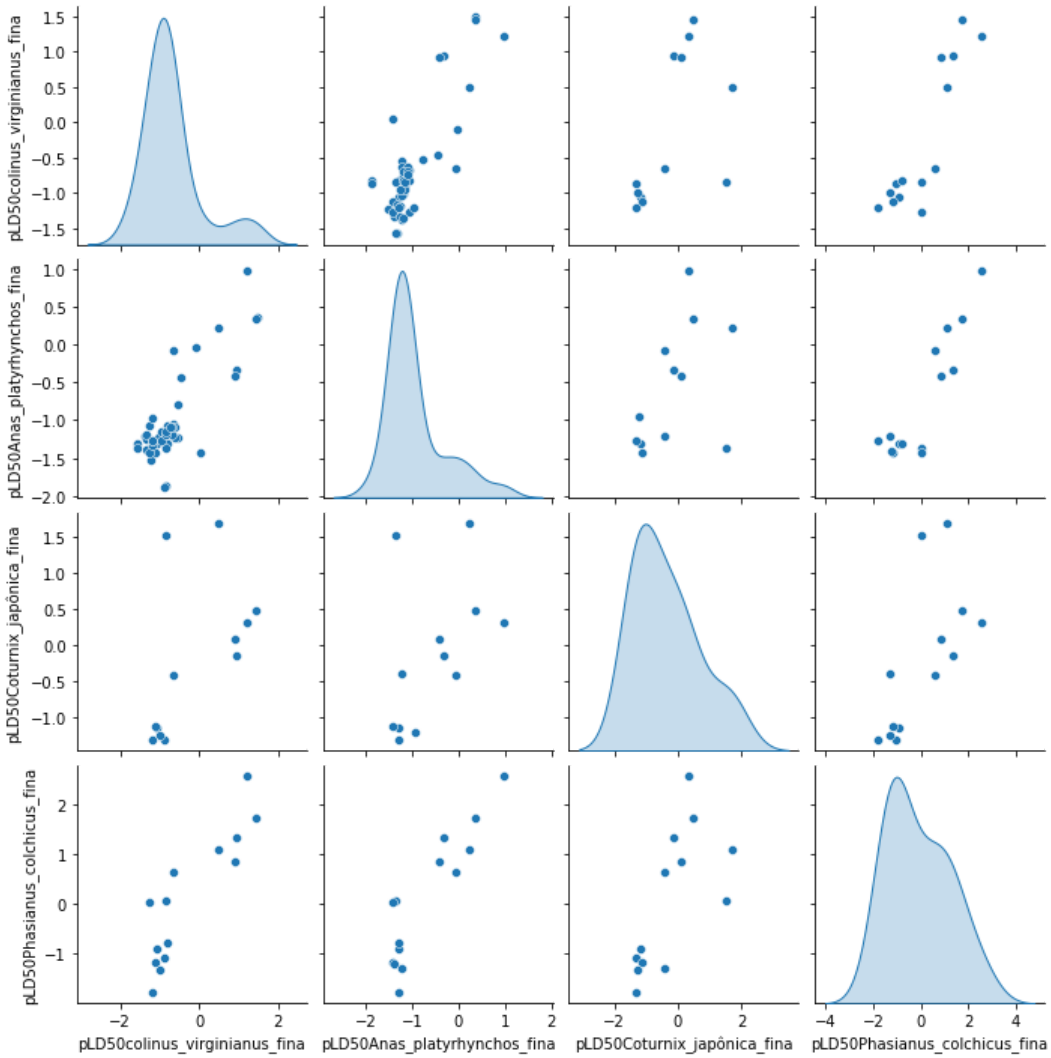



**Figura 6:** RELAÇÃO PROBABILÍSTICA - DADOS BIOLÓGICOS- CONJUNTO TESTE *(Autoria própria, 2022).*

O gráfico apresenta a distribuição dos dados conforme a faixa de valores de pLD50 entre as *tasks*. Observa-se que *A. platyrhynchos* e *C. virginianus* compartilham dados na mesma região (entre -1 e -2) e o maior desvio padrão ocorre na *task* 4 (*P. colchicus*).

Observa-se uma boa distribuição média entre os dados com um leve desvio para a direita que indica uma tendência modal. A distribuição média é mais uniforme entre as espécies que compartilham maior número de dados.

### 3.4. Modelos de Regressão

Os modelos de regressão QSAR foram desenvolvidos com hiperparâmetros otimizados utilizando redes neurais do tipo *FeedFoward*, obtendo-se os seguintes resultados:

A arquitetura da rede neural foi construída com uma camada de entrada, 7 camadas ocultas, e uma camada de saída, cada uma com o respectivo número de neurônios: 100, 75, 50, 35, 25, 15, 7, 10 e 5. Foram aplicados *drop out* de 0.3, Batch *Normalization*, algoritmo cuja função consiste em normalizar os dados tornando o aprendizado mais rápido e estável, e função de ativação LeakyReLU, que é responsável pelo processamento dos dados inseridos na rede e consequente predição da toxicidade de novos compostos conforme o conhecimento adquirido.

Adotou-se um *learning rate* (lr) = 0,1, determinando a intensidade no ajuste dos pesos neurais durante o treinamento do modelo. Utilizou-se a função de custo (*loss*) e erro global como métricas estatísticas de validação.

A adição de camadas de neurônios à rede neural aumentou a capacidade preditiva do modelo, notável pela melhoria dos resultados estatísticos. Provavelmente, devido à complexidade considerável dos dados, como observado na análise de espaço químico entre as *tasks*, o aumento de conexões entre os neurônios (devido ao aumento no número de *hidden layers*), análogo ao que ocorre no cérebro humano, pode ter contribuído para a capacidade do modelo computacional de detectar diferenças sutis entre as moléculas, que neste caso são importantes para se determinar a toxicidade de substâncias.[6]

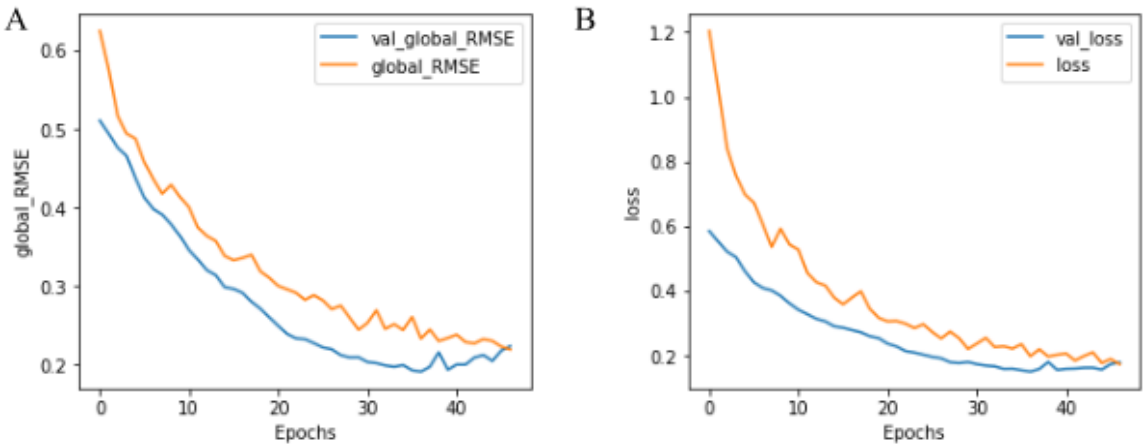


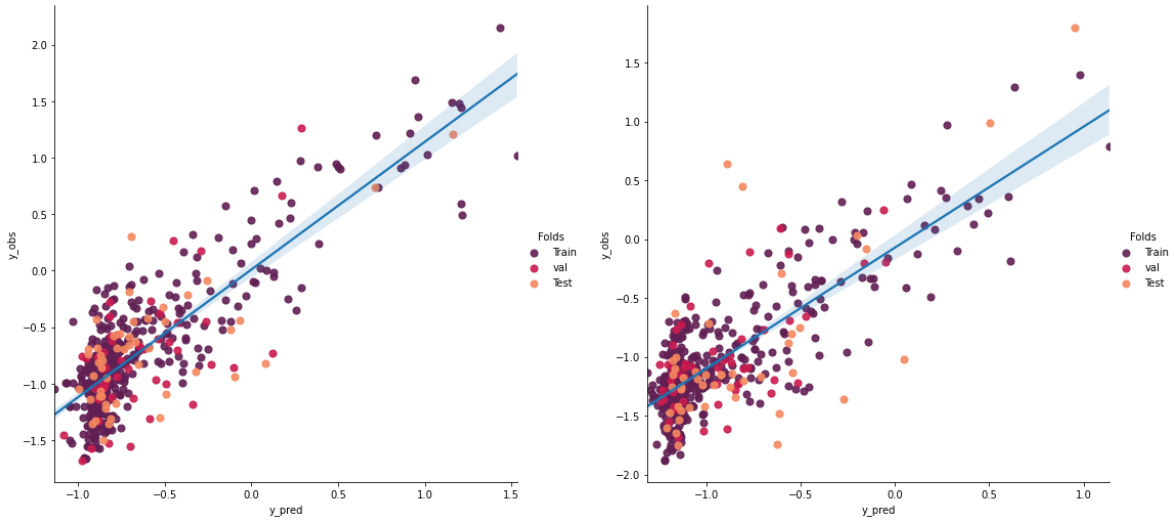

**Figura 7**: GRÁFICOS DE RMSE E FUNÇÃO *LOSS* (conjuntos treino e validação). A: erro global ( RMSE ) em função do número de épocas; B: *loss* em função do número de épocas. *(Autoria própria, 2022).*

A figura acima mostra gráficos de RMSE em função do número de épocas e função de custo (loss) em relação ao número de épocas.. Os gráficos demonstram o erro global do conjunto treino e do conjunto de validação em função do número de épocas cujo modelo é treinado. Neste caso, observa-se um erro relativamente baixo do treino e uma taxa de erro mais reduzida no conjunto validação, o que infere em uma capacidade preditiva significativa. (20)

As *tasks* 1 e 4 estão relacionadas do ponto de vista das métricas estatísticas, demonstraram melhores resultados já que os erros dos respectivos conjuntos de validação foram mais baixos, o que indica um aprendizado razoável, enquanto as *tasks* 2 e 3 apresentaram pior desempenho, com valores de erro mais elevados na validação, que associado a um baixo erro no treino é indício de uma tendência a *overffiting*. Valores mais baixos nos erros do teste também implicam em melhor performance estatística, assim como maiores valores do coeficiente de correlação de Pearson. (21)

A

B




C

D

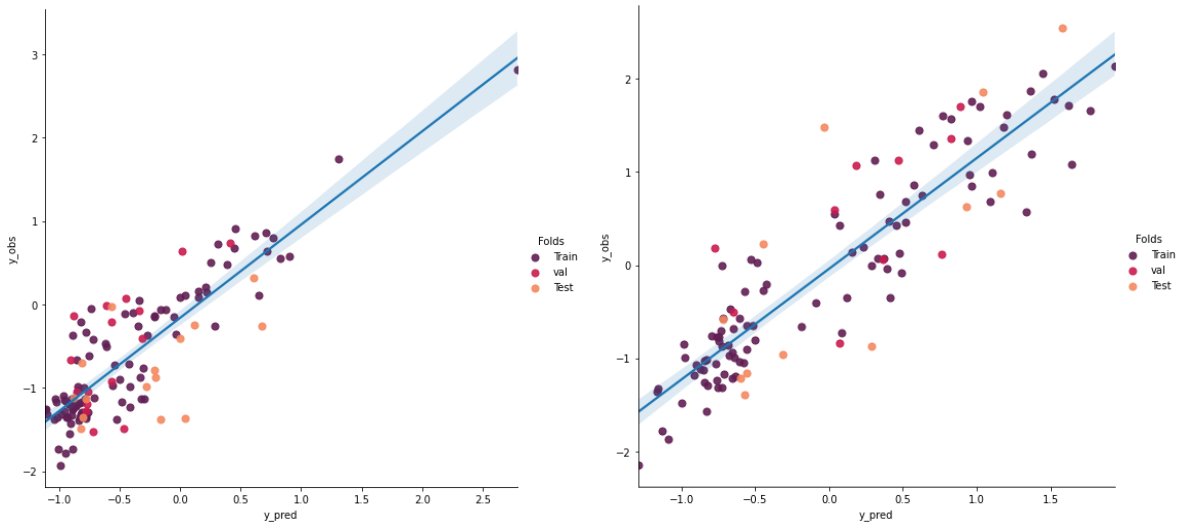



**Figura 8 :** CORRELAÇÃO ENTRE VALORES EXPERIMENTAIS E PREDITOS DE pLD50. *MULTITASK*. **A:** *C. virginianus*; **B:** *A. platyrhynchos*; **C**: *C. japonica* ; **D**: *P. colchicus.*
*(Autoria própria, 2022).*

A figura mostra a correlação entre valores observados e valores preditos de pLD50 para os conjuntos de treino, validação e teste.

Nota-se uma correlação mais relevante nas *tasks* A e D, pois apresentam pontos mais equidistantes da reta. Já nos gráficos que representam as *tasks* B e C , observa-se uma distribuição mais dispersa dos dados.

Os aglomerados de pontos com tendência de alinhamento vertical demonstram um comportamento não interessante, pois indica baixa taxa de aprendizado,ou seja, mostra que um mesmo valor predito está correspondendo a uma faixa abrangente de valores experimentais.

### 3.5. *Single Task*

#### 3.5.1. *A. platyrhynchos*

A arquitetura da rede neural do modelo ST para a *task* 1(*A. platyrhynchos)* foi construída com 5 camadas densas, cada uma com o respectivo número de neurônios: 75, 20, 25 e 5, além da camada de saída. Foram aplicados *drop out* de 0.2, *Batch Normalization* e função de ativação ‘selu’.

Adotou-se um *learning rate* (lr) = 0,1, determinando a intensidade no ajuste dos pesos neurais durante o treinamento do modelo.

Neste caso, observa-se um erro menor do treino em relação à validação, o que demonstra uma tendência do modelo a "decorar", indicando um nível de aprendizado não muito relevante.

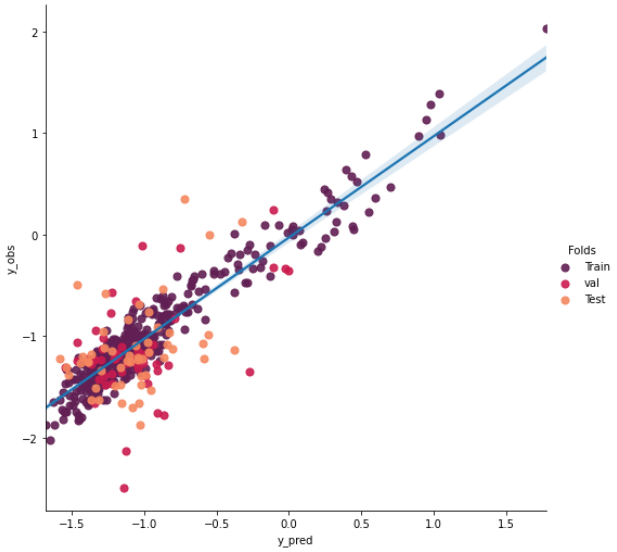



**Figura 9**: CORRELAÇÃO - VALORES EXPERIMENTAIS E PREDITOS DE pLD50. *SINGLE TASK. A. platyrhynchos. (Autoria própria, 2022).*

O gráfico acima representa a correlação entre os valores observados e valores preditos de pLD50 para os conjuntos de treino, validação e teste da *task* 1 (*A. platyrhynchos*). Os aglomerados de pontos com tendência de alinhamento vertical demonstram um comportamento não interessante, pois indica taxa de aprendizado menor .

### *3.5.2. C. virginianus*

A arquitetura da rede neural do modelo ST para a *task* 2 (*C. virginianus)* foi construída com 4 camadas densas, cada uma com o respectivo número de neurônios: 20, 15 e 10, além da camada de saída. Foram aplicados *drop out* de 0.3, Batch Normalization e função de ativação 'relu'.

Neste caso, também se observa um erro relativamente alto no conjunto de validação em relação ao conjunto de treino, o que demonstra uma tendência ao *overfitting*, que inclusive foi maior nesta *task* 2 em comparação à *task* 1.

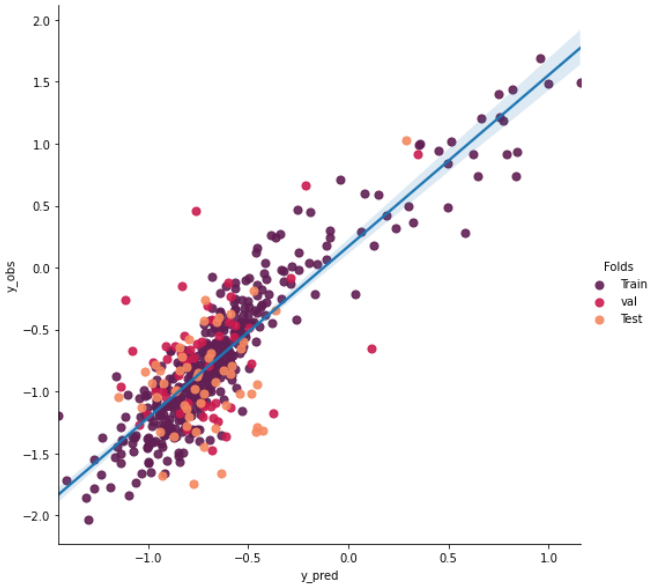



**Figura 10** : CORRELAÇÃO - VALORES EXPERIMENTAIS E PREDITOS DE pLD50. *SINGLE TASK*. *C. virginianus*. *(Autoria própria, 2022).*

O gráfico acima representa a correlação entre os valores observados e valores preditos de $pLD_{50}$ para os conjuntos de treino, validação e teste da task 2 (*C. virginianus*). Nota-se uma correlação levemente mais relevante nesta *task*, pois apresentam pontos mais próximos de serem equidistantes da reta.

### *3.5.3. C. japônica*

A arquitetura da rede neural do modelo ST para a *task* 3 (*C. japonica* ) foi construída com 4 camadas densas, cada uma com o respectivo número de neurônios: 60, 40 e 5, além da camada de saída. Foram aplicados *drop out* de 0.2, *Batch Normalization* e função de ativação 'selu'.

Neste caso, também se observa um erro bastante alto no conjunto de validação e de teste em relação ao conjunto de treino, o que demonstra uma tendência a *overfitting,* que inclusive foi a maior entre todas as *tasks*.

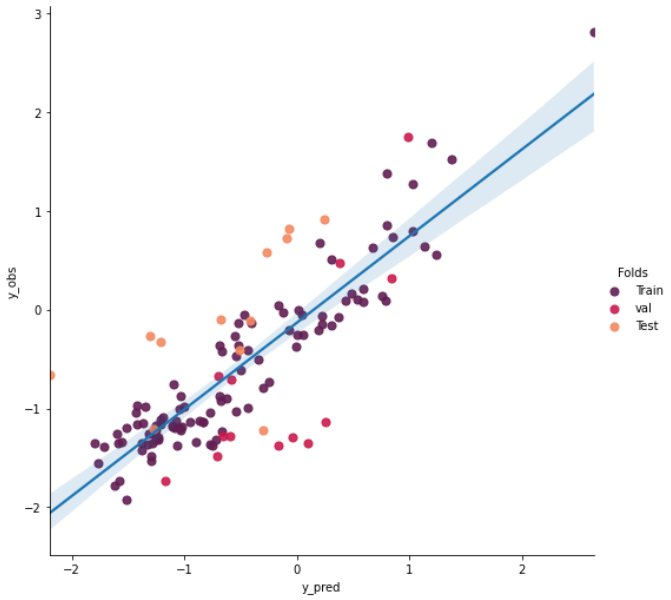



**Figura 11:** CORRELAÇÃO - VALORES EXPERIMENTAIS E PREDITOS DE pLD50. *SINGLE TASK C. japônica. (Autoria própria, 2022).*

O gráfico acima representa a correlação os valores observados e valores preditos de pLD50 para os conjuntos de treino, validação e teste da task 3 (*C. japônica.*). Nota-se uma dispersão considerável dos dados, ou seja, uma baixa correlação entre os mesmos ,o que implica em baixa performance de aprendizado para esta *task*, que apresentou métricas estatísticas inferiores às demais.

### *3.5.4. P. colchicus*

A arquitetura da rede neural do modelo ST para a *task* 4 (*P. colchicus*) foi construída com 4 camadas densas, cada uma com o respectivo número de neurônios: 50, 40 e 5, além da camada de saída. Foram aplicados *drop out* de 0.2, *Batch Normalization* e função de ativação 'selu'.

Neste caso, também se observa um erro relativamente alto no conjunto de validação e teste em relação ao conjunto de treino, o que demonstra uma tendência a *overffiting*, semelhante à *task* 3.

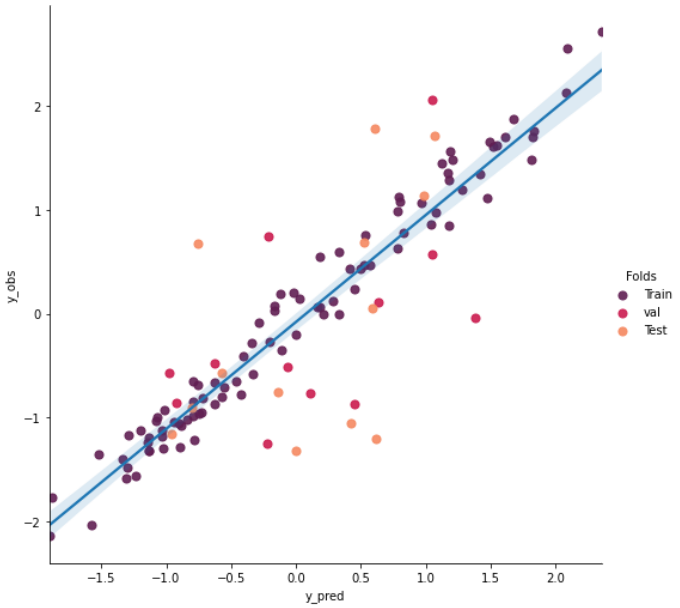



**Figura 12:** CORRELAÇÃO - VALORES EXPERIMENTAIS E PREDITOS DE pLD50. *SINGLE TASK. P. colchicus.* *(Autoria própria, 2022).*

A figura acima representa a correlação entre os valores observados e valores preditos de pLD50 para os conjuntos de treino, validação e teste da *task* 4 (*P. colchicus*). Nota-se uma correlação um pouco mais relevante entre os dados desta *task*, assim como uma melhor distribuição dos dados, mas ainda assim não demonstrou taxa de aprendizado muito significante.

### 3.6. Modelos de Regressão *Multi task x Single task*

Os resultados *Multitask* (MT) demonstraram melhor desempenho comparando-se com os resultados dos modelos *Single task*. O que corrobora com a proposta da abordagem MT que consiste em minimizar o custo de treinamento de todas as *tasks* e melhorar a capacidade de performance das predições. [26] Neste caso, portanto, visa predizer a atividade toxicológica de compostos químicos para diferentes alvos.

As tabelas a seguir apresentam os valores obtidos das métricas estatísticas e mostram como o aprendizado MT permite explorar o compartilhamento de informação química e biológica entre as espécies alvo, o que justifica seu melhor desempenho preditivo.

**Tabela 2:** MÉTRICAS ESTATÍSTICAS – DADOS CONJUNTO TREINO. Coeficiente de *Pearson* (correlação); erro absoluto médio (MAE); erro quadrático médio (MSE); raiz quadrada do erro médio (RMSE).

| | task | r Pearson | RMSE | MSE | MAE |
|---|---|---|---|---|---|
| **MultiTask** | *A.platyrhynchos* | 0,83 | 0,4 | 0,16 | 0,31 |
| | *C.virginianus* | 0,82 | 0,38 | 0,14 | 0,27 |
| | *C. japônica* | 0,86 | 0,53 | 0,28 | 0,41 |
| | *P. colchicus* | 0,91 | 0,47 | 0,23 | 0,38 |
| **SingleTask** | *A.platyrhynchos* | 0,92 | 0,24 | 0,19 | 0,18 |
| | *C.virginianus* | 0,92 | 0,27 | 0,08 | 0,22 |
| | *C. japônica* | 0,96 | 0,25 | 0,07 | 0,20 |
| | *P.colchicus* | 0,97 | 0,27 | 0,08 | 0,21 |

Observa-se no gráfico acima que no conjunto treino a correlação foi mais alta no modelo ST em comparação ao MT para todas as *tasks*. Já os índices de erro foram mais elevados no modelo MT em relação ao ST. Esta alta taxa de correlação e baixos valores de erro no ST indicam uma possível tendência a *overffiting*[13], que implica em um indicativo de aprendizado menor em relação ao MT.

**Tabela 3:** MÉTRICAS ESTATÍSTICAS – DADOS CONJUNTO VALIDAÇÃO. Coeficiente de *Pearson* (correlação); erro absoluto médio (MAE); erro quadrático médio (MSE); raiz quadrada do erro médio (RMSE).

| | task | r Pearson | RMSE | MSE | MAE |
|---|---|---|---|---|---|
| **MultiTask** | *A.platyrhynchos* | 0,70 | 0,54 | 0,29 | 0,41 |
| | *C.virginianus* | 0,53 | 0,41 | 0,16 | 0,31 |
| | *C. japônica* | 0,62 | 0,68 | 0,47 | 0,54 |
| | *P. colchicus* | 0,52 | 0,96 | 0,93 | 0,83 |
| **SingleTask** | *A.platyrhynchos* | 0,45 | 0,43 | 0,19 | 0,32 |
| | *C.virginianus* | 0,53 | 0,57 | 0,32 | 0,44 |
| | *C. japônica* | 0,37 | 0,83 | 0,69 | 0,70 |
| | *P. colchicus* | 0,62 | 0,85 | 0,72 | 0,70 |

Avaliando-se, em seguida, os resultados obtidos pelo conjunto validação, nota-se valores elevados dos índices de erro, inclusive maiores que o índice de correlação, no caso das *tasks* 3 e 4 (*C. japônica e P. colchicus,* respectivamente). Esses resultados, corroboram com a análise dos resultados do conjunto treino e indicam um desempenho menor destas *tasks* em relação às demais. Nestas, os índices de erro foram menores em comparação ao r pearson , demonstrando um melhor desempenho

**Tabela 4:** MÉTRICAS ESTATÍSTICAS – DADOS CONJUNTO TESTE. Coeficiente de *Pearson* (correlação); erro absoluto médio (MAE); erro quadrático médio (MSE); raiz quadrada do erro médio (RMSE).

| | task | r Pearson | RMSE | MSE | MAE |
|---|---|---|---|---|---|
| **MultiTask** | *A.platyrhynchos* | 0,70 | 0,38 | 0,14 | 0,29 |
| | *C.virginianus* | 0,71 | 0,53 | 0,28 | 0,41 |
| | *C. japônica* | 0,59 | 0,69 | 0,48 | 0,60 |
| | *P. colchicus* | 0,80 | 0,80 | 0,65 | 0,72 |
| **SingleTask** | *A.platyrhynchos* | 0,46 | 0,5 | 0,26 | 0,39 |
| | *C.virginianus* | 0,63 | 0,49 | 0,24 | 0,39 |
| | *C. japônica* | 0,25 | 0,95 | 0,92 | 0,81 |
| | *P. colchicus* | 0,57 | 0,82 | 0,67 | 0,70 |

De modo semelhante aos resultados obtidos pelo conjunto validação, nota-se, nos resultados do conjunto teste, valores elevados dos índices de erro, maiores que o índice de correlação, no caso das tasks 3 e 4 (*C. japônica e P. colchicus,* respectivamente), sendo que o ST para *C. japônica apresentou* as piores métricas. As *tasks* 1 e 2 (*A.platyrhynchos e C.virginianus* ) apresentaram índices de erro menores e maiores valores de correlação de Pearson, demonstrando um melhor desempenho, que correspondem às *tasks* que compartilham maior número de dados.

## 4. CONCLUSÃO

A partir deste trabalho pode-se concluir que a ecotoxicologia computacional demonstra potencial na predição de toxicidade e avaliação de risco de compostos químicos, contribuindo com ferramentas consideráveis para orientar legislações que regulam a utilização de agrotóxicos e pesticidas, com intuito de evitar o uso de produtos químicos potencialmente agressivos ao meio ambiente, além de serem capazes de auxiliar no desenvolvimento de novos insumos químicos menos tóxicos. A ecotoxicologia computacional demonstra ser uma alternativa promissora em relação a metodologias tradicionais que ainda realizam testes em

animais. Todavia, é importante ressaltar que ainda se tratam de ferramentas contribuintes e não totalmente substitutas de outros métodos de avaliação de toxicidade. Nota-se, também, que assim como outras metodologias, as ferramentas preditivas computacionais apresentam suas falhas e também precisam ser aprimoradas, para que se obtenham resultados melhores e se tornem ainda mais promissoras.

## 5. PERSPECTIVAS

As perspectivas em relação a este trabalho consistem em aplicar ferramentas de *Transfer learning,* e explorar outras metodologias de machine learning para realização de *baseline*, e realizar interpretação mecanística dos modelos com objetivo de melhorar a performance preditiva dos modelos e sua interpretabilidade.

## 6. REFERÊNCIAS BLBLIOGRÁFICAS


(1) Pignati, W. A., Lima, S. N. A. F., Lara, S. S., Correa, M. L. M., Barbosa, R. J., Leão, C. H. L. & Pignatti, G. M. **2017**. Distribuição espacial do uso de agrotóxicos no Brasil: uma ferramenta para a Vigilância em Saúde. Ciência & Saúde Coletiva. 22(10), 3281-3293.

(2) Parker, M. L., & Goldstein, M. I. (**2000**). Differential toxicities of organophosphate and carbamate insecticides in the nestiling Eutopean starling (Sturnus vulgaris). Archives of environmental contamination and toxicology, New York, 39(2), 233-242.

(3) Amâncio, S., Souza, V. B., & Melo, C. (**2008**). Columba livia e Pitangus sulphuratus como indicadoras de qualidade ambiental em área urbana. Revista Brasileira de Ornitologia. 16(1), 32-7.

(4) [OECD] Organisation for Economic Co-operation and Development. 2010a. Test No. 223: Avian acute oral toxicity test. Paris (FR). [cited 2016 August 26]. https://doi.org/10.1787/9789264090897-en

(5) Fernando D. Prieto-Martínez Edgar López-López K. Eurídice Juárez-Mercado José L. Medina-Franco, in Silico Drug Design, **2019** .

(6) Schmidhuber, Jürgen."Aprendizagem profunda em redes neurais: uma visão geral". Redes Neurais . 61 : 85-117. arXiv : 1404.7828 . doi : 10.1016/j.neunet.**2014**.09.003

(7) Jiarui Chen, Hong Hin Cheong, and Shirley W. I. Siu Journal of Chemical Information and Modeling **2021** 61 (8), 3789-3803.

(8) EPA. Ecotoxicology Database (ECOTOX) https://cfpub.epa.gov/ecotox/ (accessed May 18, **2021**)

(9) OpenFoodTox: EFSA's Open Source Toxicological Database on Chemical Hazards in Food and Feed. https://doi.org/10.2903/j.efsa.**2017**.e15011

(10) Vinicius M. Alves, Rodolpho C. Braga, Eugene N. Muratov e Carolina Horta Andrade. QUIMIOINFORMÁTICA: UMA INTRODUÇÃO .Quim. Nova, Vol. 41, No. 2, 202-212, **2018**.

(11) FOURCHES, D ;et al. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J. Chem. Inf. Model. **2010**, 50 (7), 1189–1204.

(12) GraphPadPrism .https://www.graphpad.com/scientific-software/prism.**2021**

(13) KAR, S; et al. Is intraspecies QSTR model answer to toxicity data gap filling: Ecotoxicity modeling of chemicals to avian species. Science of The Total Environment,Vol.738,139858 (2020).https://doi.org/10.1016/j.scitotenv**.2020**.139858

(14) L.J.P.Maaten and G.E. Hinton. **Visualizing High-Dimensional Data Using t-SNE.** *Journal of Machine Learning Research* 9(Nov):2579-2605, **2008**.

(15) RINIKER, S; et al. Similarity Maps - A Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. J. Cheminform. **2013**, 5 (9), 1–7

(16) Python v3.10.6. https://www.python.org/. 2021

(17) Kevin Greenman (**2022**), "Message-Passing Neural Networks for Molecular Property Prediction Using Chemprop," https://nanohub.org/resources/36082.

(18) ABADI, M; et al. TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A System for Large-Scale Machine Learning. In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16); **2016**; pp 265–284.

(19) CONNSONNI, V ; et al. Comments on the Definition of the Q 2 Parameter for QSAR Validation. J. Chem. Inf. Model**. 2009**, 49 (7), 1669–1678

(20) *McInnes, L, Healy, J,* UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426, 2018*

(21) LECUN, Y; et al. Deep Learning. Nature **2015**, 521 (7553), 436–444.