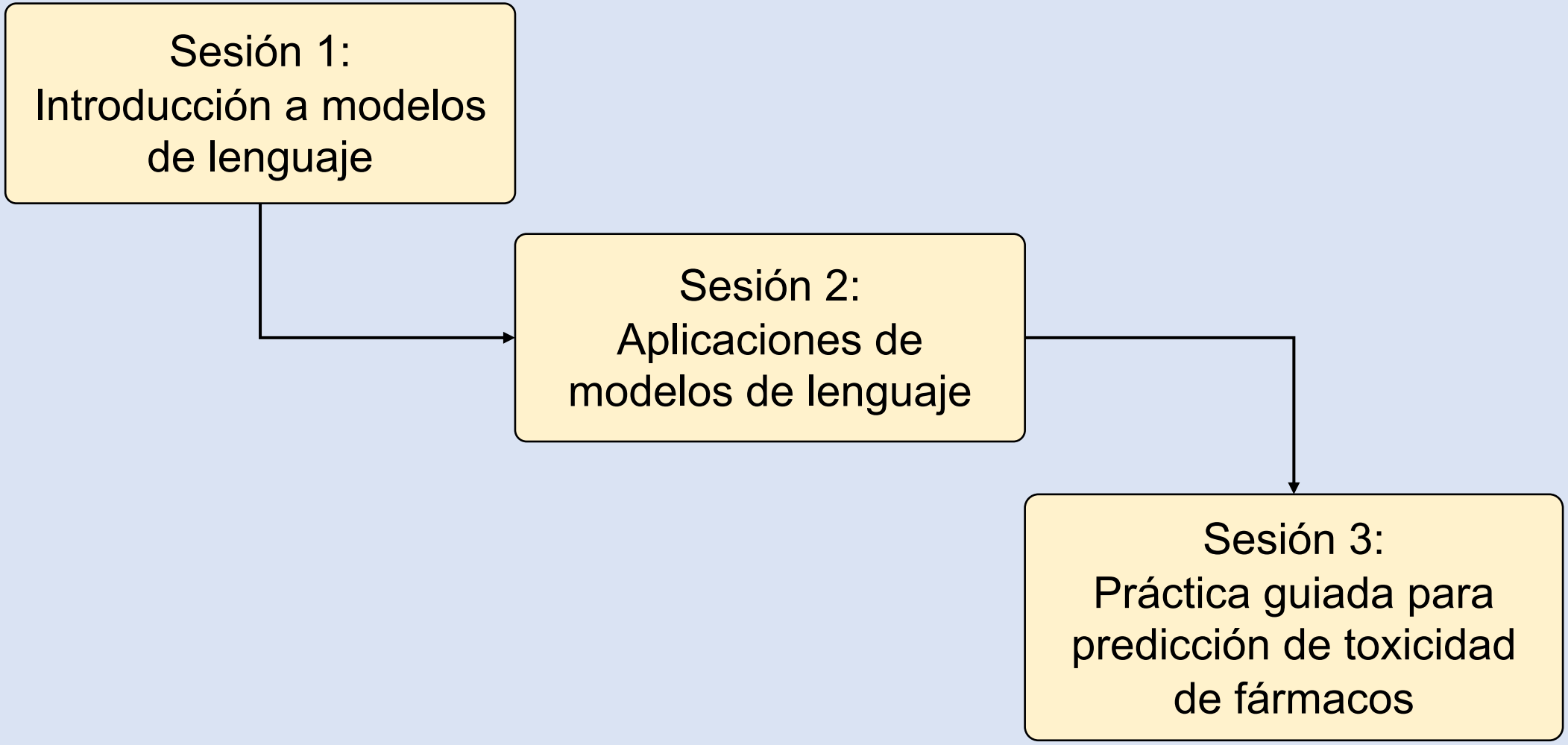
Modelos que aprenden el lenguaje de las moléculas

Raúl Fernández Díaz
Estudiante de doctorado industrial

University College Dublin
University for All

Organización



Sesión 2: Contenidos

¿Cómo podemos utilizar los modelos de lenguaje molecular?

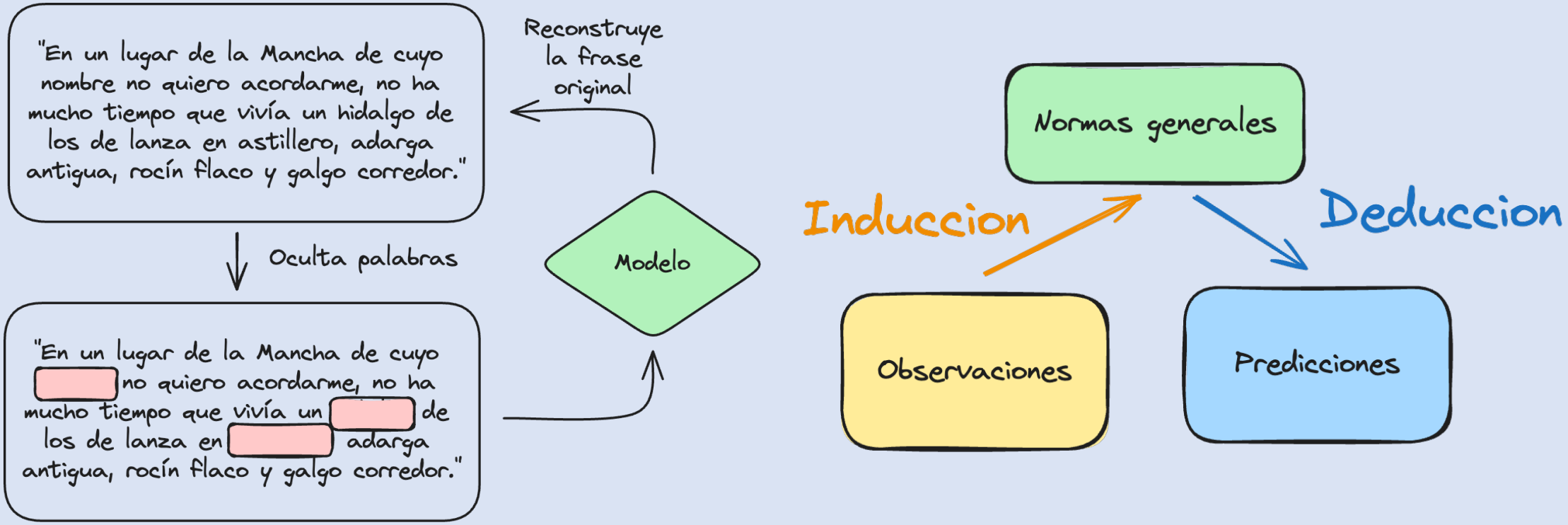
¿Cómo podemos evaluarlos?

Principales herramientas

Cuarta Parte

¿Cómo podemos utilizar los modelos de lenguaje molecular?

Transferencia de conocimiento

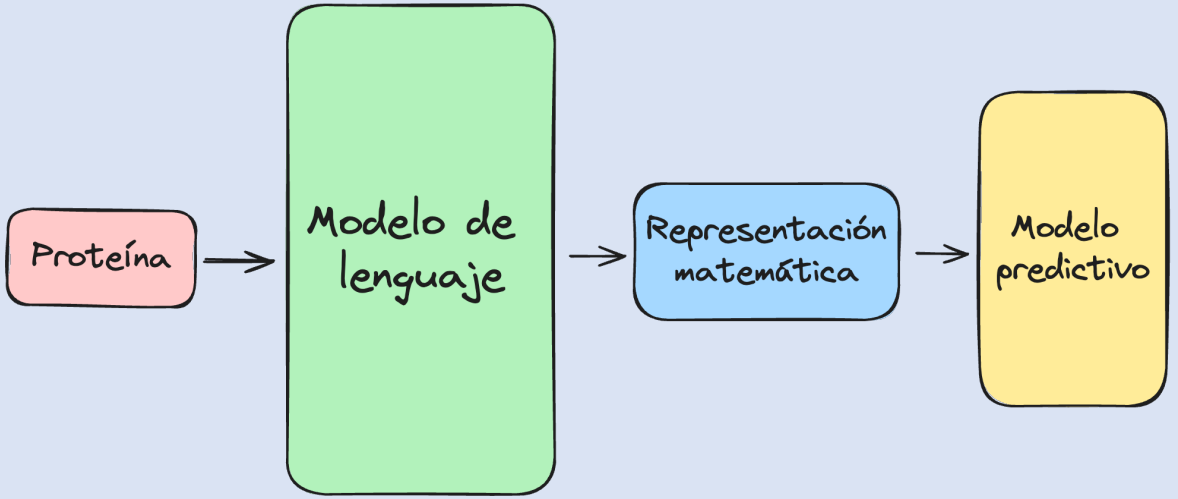


Transferencia

Transferencia de conocimiento: estrategias

Transferencia de representaciones:

* El modelo no cambia

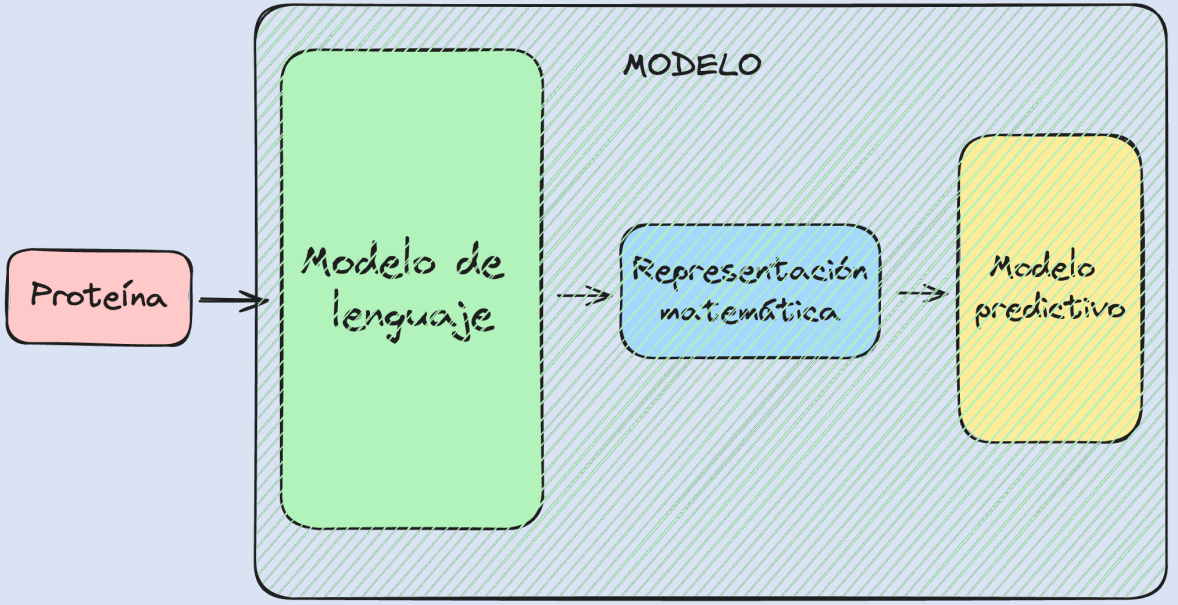


Ventajas: Muy eficiente

Desventaja: Utilidad limitada

Transferencia de aprendizaje:

* El modelo cambia



Ventajas: No es eficiente

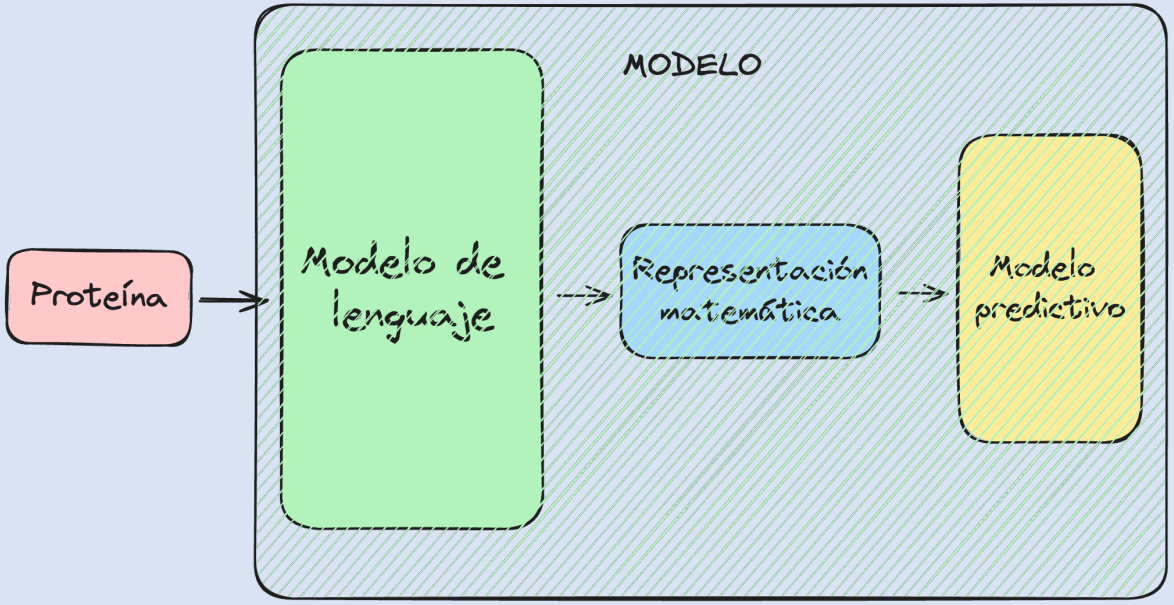
Desventaja: Mucho más versátil

Transferencia

Transferencia de conocimiento: estrategias

Transferencia de aprendizaje:

* El modelo cambia

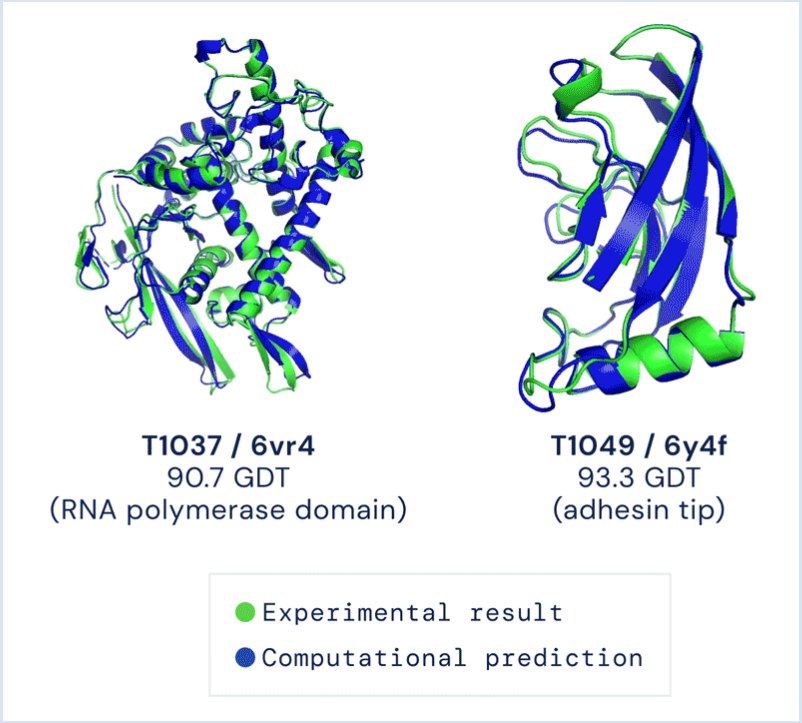


Ventajas: No es eficiente

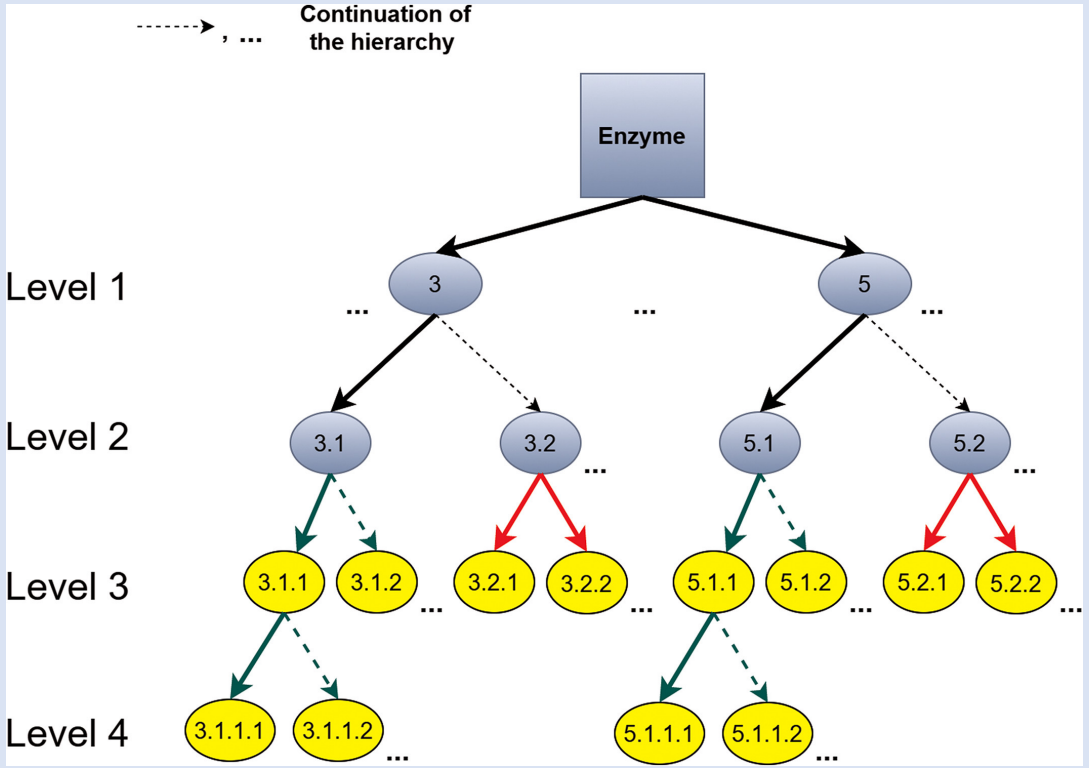
Desventaja: Mucho más versátil

Transferencia de conocimiento: aplicaciones

Predicción de la estructura de proteínas

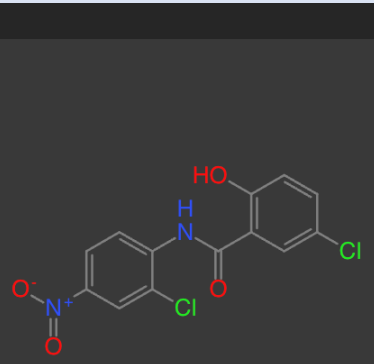


Clasificación de enzimas



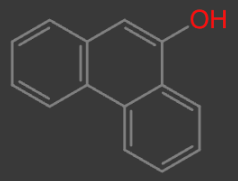
Transferencia

Transferencia de conocimiento: aplicaciones



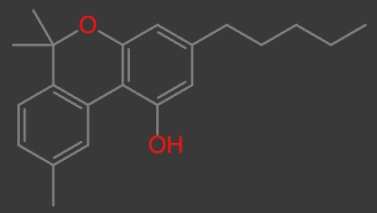
Tox 21 avrg: 0.268
- **NR-AR**: 0.009
- **NR-AR-LBD**: 0.008
- **NR-AhR**: 0.635
- **NR-Aromatase**: 0.150
- **NR-ER**: 0.109
- **NR-ER-LBD**: 0.069
- **NR-PPAR-gamma**: 0.021
- **SR-ARE**: 0.717
- **SR-ATAD5**: 0.013
- **SR-HSE**: 0.267
- **SR-MMP**: 0.984
- **SR-p53**: 0.234

ESOL: −0.781
HIV: 0.013
Smiles: **OC1=C(C=C(Cl)C=C1)C(...**



Tox 21 avrg: 0.381
- **NR-AR**: 0.020
- **NR-AR-LBD**: 0.009
- **NR-AhR**: 0.566
- **NR-Aromatase**: 0.189
- **NR-ER**: 0.923
- **NR-ER-LBD**: 0.763
- **NR-PPAR-gamma**: 0.036
- **SR-ARE**: 0.849
- **SR-ATAD5**: 0.047
- **SR-HSE**: 0.083
- **SR-MMP**: 0.978
- **SR-p53**: 0.114

ESOL: −0.598
HIV: 0.004
Smiles: **OC1=CC2=C(C=CC=C2)...**



Tox 21 avrg: 0.420
- **NR-AR**: 0.007
- **NR-AR-LBD**: 0.008
- **NR-AhR**: 0.767
- **NR-Aromatase**: 0.667
- **NR-ER**: 0.601
- **NR-ER-LBD**: 0.326
- **NR-PPAR-gamma**: 0.033
- **SR-ARE**: 0.822
- **SR-ATAD5**: 0.069
- **SR-HSE**: 0.325
- **SR-MMP**: 0.990
- **SR-p53**: 0.422

ESOL: −1.282
HIV: 0.004
Smiles: **CCCCCC1=CC2=C(C(O)...**

MolFormer-XL

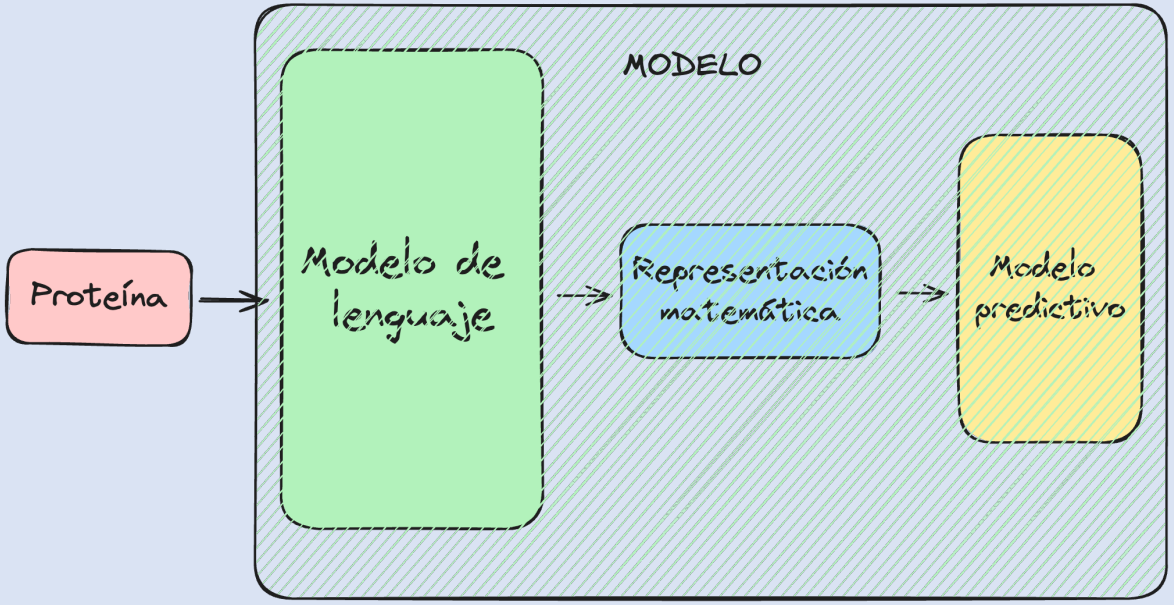
https://molformer.res.ibm.com/

Transferencia

Transferencia de conocimiento: estrategias

Transferencia de aprendizaje:

* El modelo cambia



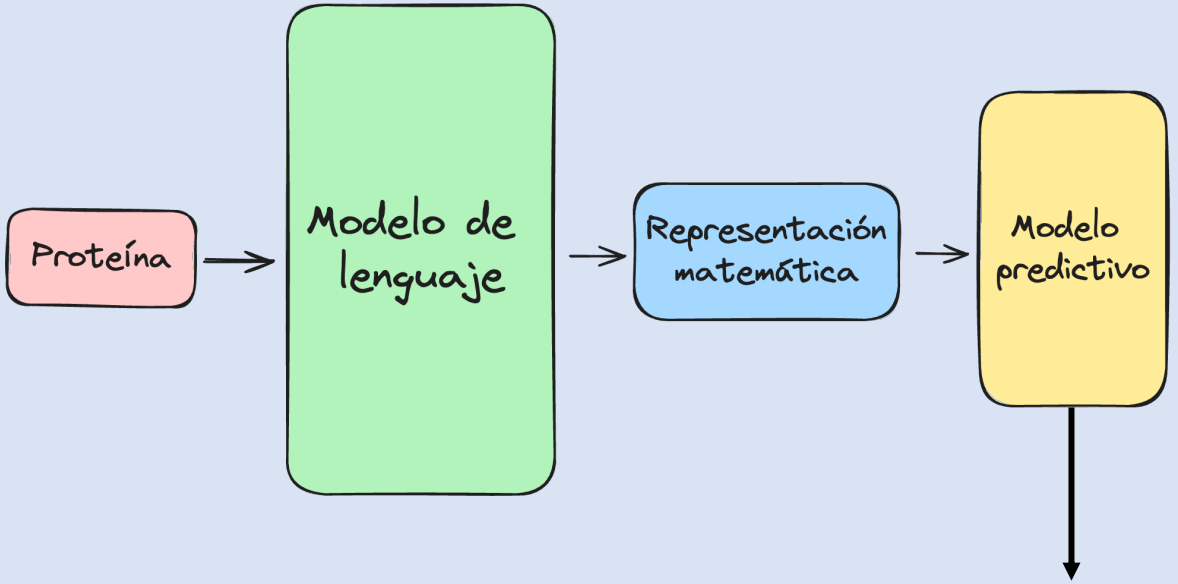
Ventajas: No es eficiente

Desventaja: Mucho más versátil

Transferencia de conocimiento: estrategias

Transferencia de representaciones:

* El modelo no cambia



Ventajas: Muy eficiente

Desventaja: Utilidad limitada

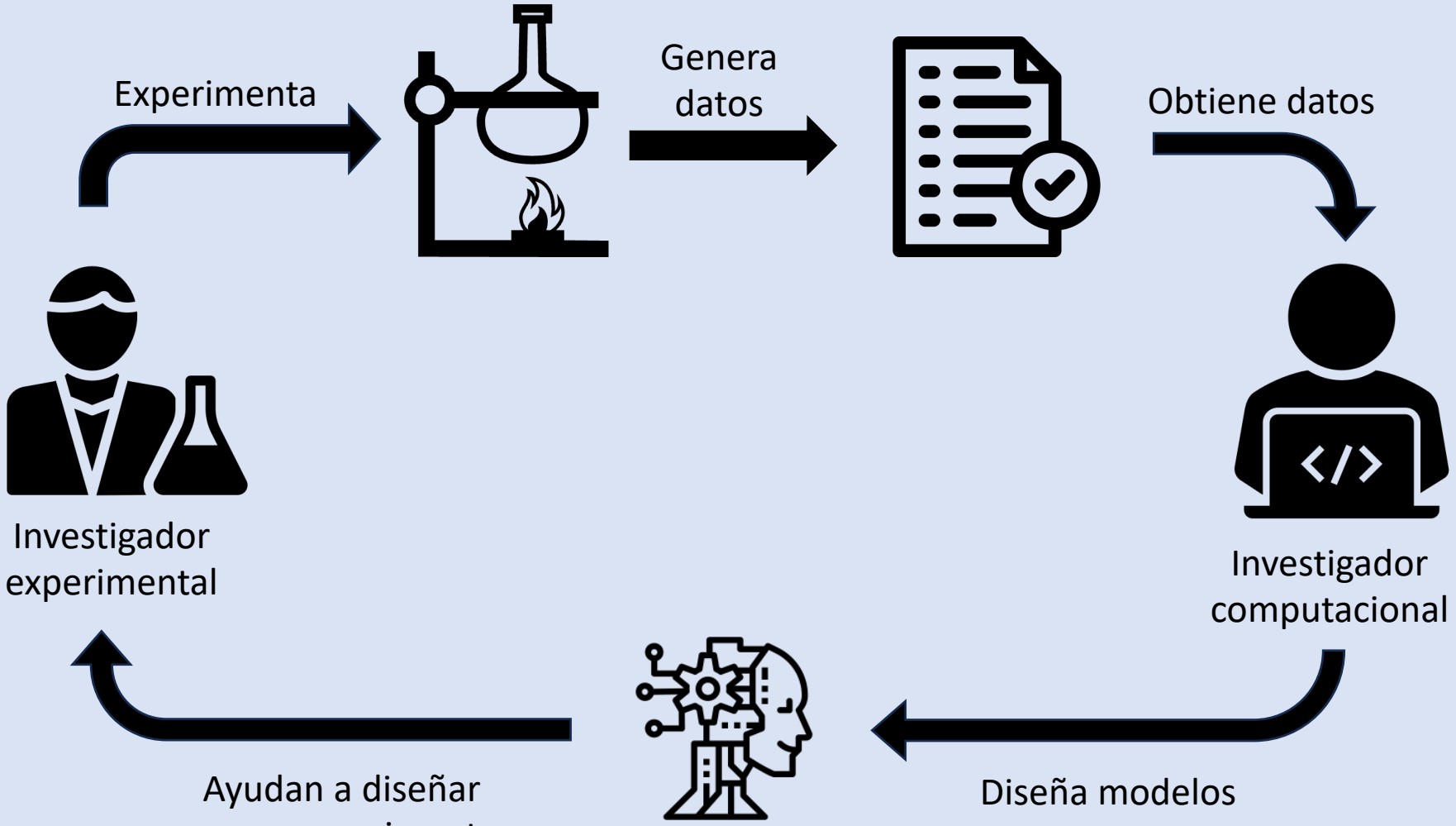
ML tradicional

- Support Vector Machines
- Random Forest
- K-Nearest Neighbours
- Gradient boosting (extreme, adaptive, or light)
- Ensemble

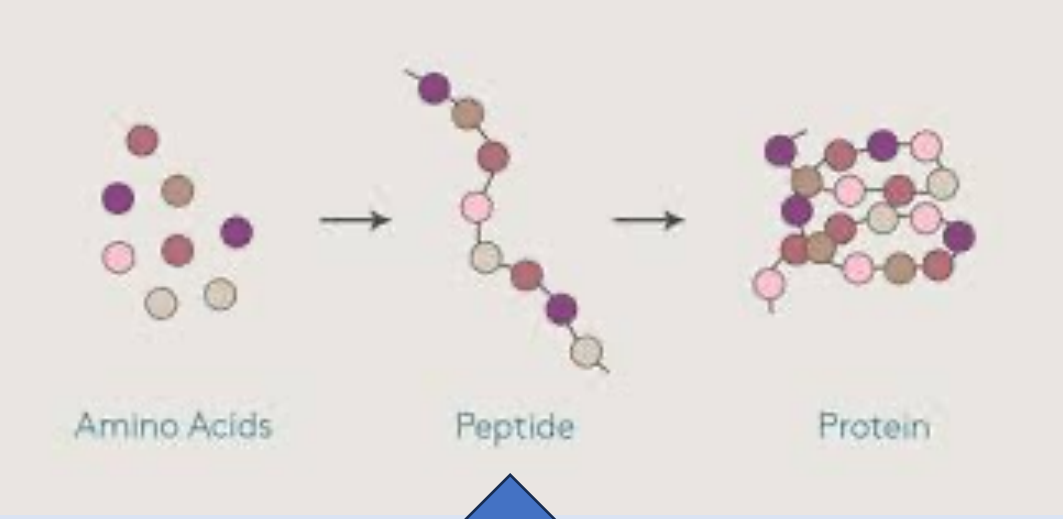
Métodos neurales (deep learning)

- Multi-layer perceptron
- Redes neurales convolucionales
- Redes neurales recursivas
- Transformers

Transferencia de representaciones: Automatización



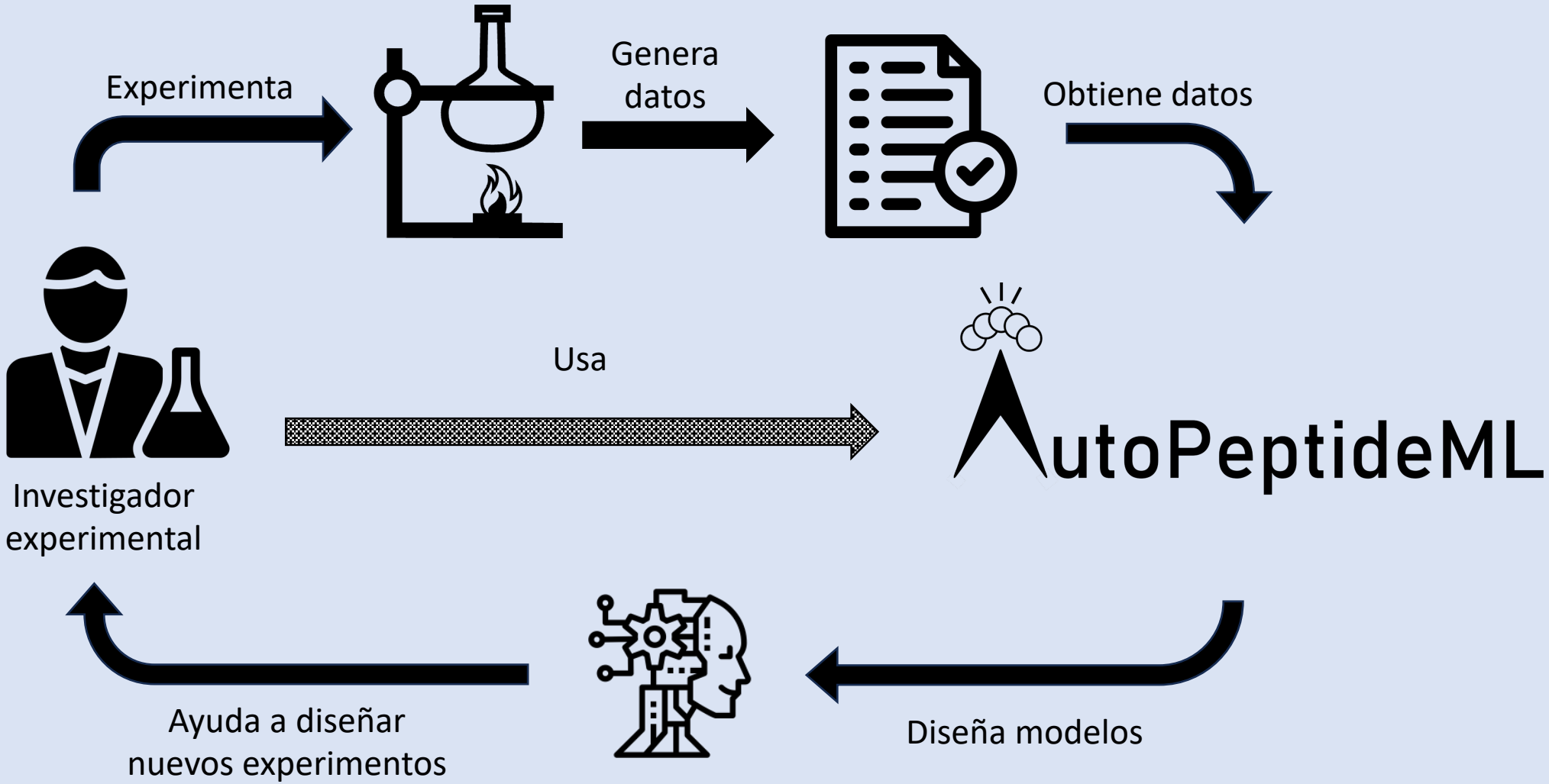
Transferencia de representaciones: AutoPeptideML



Amino acid sequence	Therapeutic application (bioactivity)*
α_{s2}-casein f(203–208)	Antimicrobial; antihypertensive; antioxidant
Ile-Pro-Pro; Val-Pro-Pro	Antihypertensive*
(Tyr-Pro-Phe-Pro-Gly-Pro-Ile-Pro-Asn-Ser-Leu) β-casein f(60–70)	Immunostimulatory, opioid agonist; ACE-inhibitory
Unidentified	Immunostimulatory, ACE-inhibitory
Arg-Val-Pro-Ser-Leu	ACE-inhibitory
Phe-Arg-Asp-Glu-His-Lys-Lys; and Lys-His-Asp-Arg-Gly-Asp-Glu-Phe	Antioxidative
Ile-Thr-Pro; Ile-Ile-Pro; Gly-Gln-Tyr; Ser-Thr-Tyr-Gln-Thr	ACE-inhibitory

Transferencia

Transferencia de representaciones: AutoPeptideML



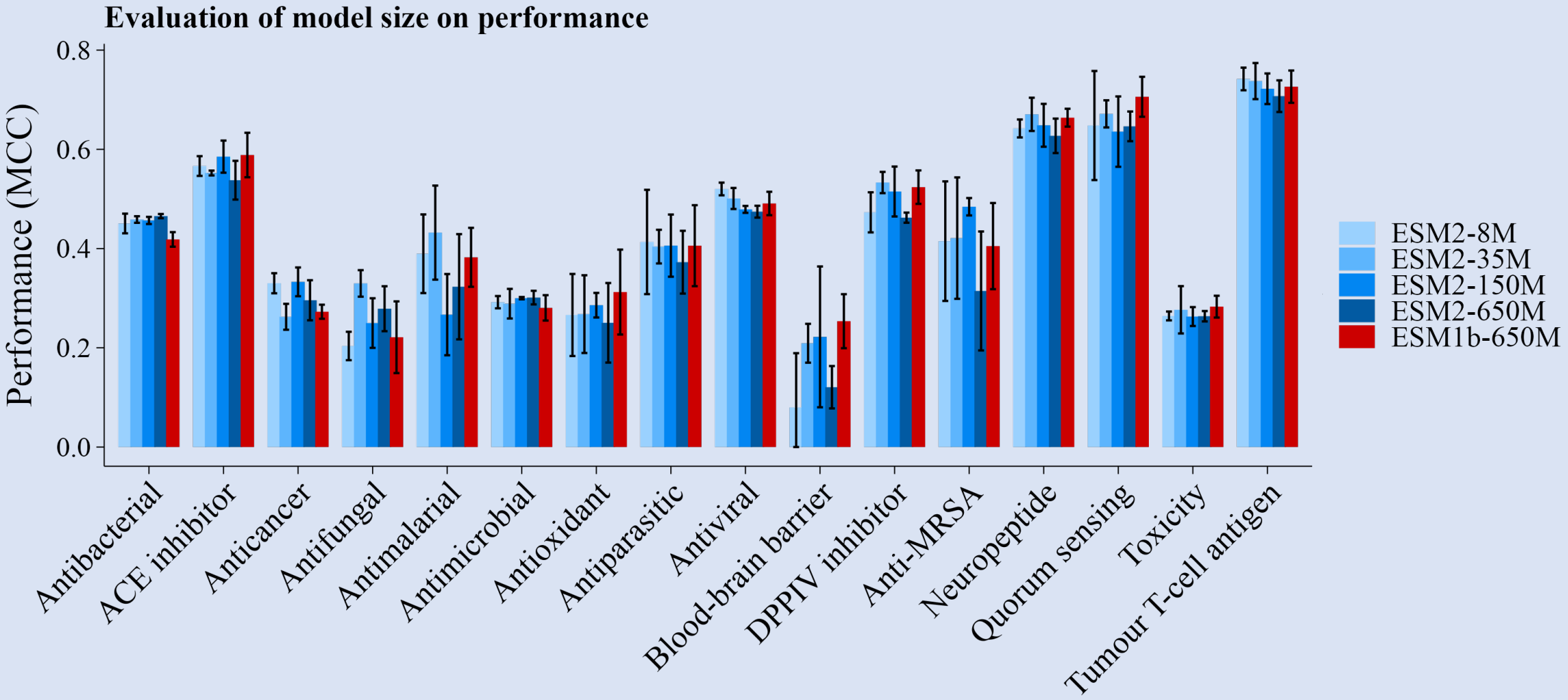
Transferencia

Transferencia de representaciones: AutoPeptideML

AutoPeptideML

http://peptide.ucd.ie/AutoPeptideML

Transferencia de representaciones: AutoPeptideML



Transferencia

Transferencia de representaciones: Automatización

Transferencia

Proteínas

Feature Types	Description	References
Amino Acid Composition	Simplest, primary, and fundamental	8,9,22
Sequence Order	Capture all possible combinations of amino acids in oligomeric proteins, exceptionally large number of features	40–45
Physicochemical Properties	Classify amino acids based on properties; Composition, order, and position-specific information are usually extracted	1,36,46
Multiprofile Bayes	Incorporate both position-specific information and the posterior probability of each amino acid type	16,52,53
Secondary Structure Based Features	Classify amino acids according to their tendency to form a specific secondary structural element	1,23,25,48
PSSM-based Probability	Evolutionary information was included by a position-specific scoring matrix	16,18,19,54
Fourier Transform Based Feature	Extract low frequency coefficients in frequency domain	15,27,55,57
Functional Domain Composition	Convert protein sequence into a sequence of functional domain types	37
Split Amino Acid Composition	Incorporate both position-specific information and amino acid composition	16

Moléculas pequeñas (fármacos)

Descriptor type	Descriptors calculated
Constitutional	Molecular Weight (*MW*), Rotational Bonds (*RotB*), Hydrogen Bond Acceptors (*HBA*) and Hydrogen Bond Donors (*HBD*)
Thermodynamic	Heat of Formation (*HF*), Log of Partition Coefficient (*LogP*), Standard Gibbs Free Energy (*G*), Stretch Energy (*Es*), Torsional Energy (*Et*) and Total Energy (*E*)
Electronic	Dipole (*DPL*), Electronic Energy (*ElecE*) and Molecular Polar Surface Area (*MPSA*)
Steric and/or Spatial	Molar Refractivity (*MR*) and Molar Volume (*MV*)
Topological	Balaban Index (*BIndx*), Cluster Count (*ClsC*), Shape Coefficient (*ShpC*), Total Connectivity (*TCon*), Molecular Topological Index (*TIndx*) and Wiener Index (*WIndx*)
Semi-empirical (Quantum chemical)	Energy of Highest Occupied Molecular Orbital (*HOMOEnergy*) and Energy of Lowest Unoccupied Molecular Orbital (*LUMOEnergy*)