UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI CHIMICA

Doctoral School in Industrial Chemistry
XXXIV Cycle

Molecular Dynamics and Cheminformatics Methods to Explore the Chemical Reality

Enrico Gandini

Tutor: Prof. Stefano Pieraccini
Cotutor: Prof. Daniele Passarella
Coordinator: Prof. Dominique Marie Roberto

Academic Year 2021/2022

Non si può mica pretendere che un chimico sappia il mondo a mente.
(Italo Svevo, *La coscienza di Zeno*)

Contents

List of Figures

List of Tables

Acronyms

Antifreeze Protein (AFP) A class of structurally diverse proteins that protect different species of living organisms from fatally freezing in icy environments. The function of AFPs is similar, but their sequence, structure, and efficacy varies enormously. iv, 7–10

Area Under the Receiver Operating Characteristic curve (AUROC) A Receiver Operating Characteristic (ROC) curve is a graphical plot which illustrates the performance of a classification model as its discrimination threshold is varied. By computing the Area Under the ROC curve (AUROC), the information is summarized in one number. For a binary classification problem, the AUROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. AUROC is between 0 and 1, an higher value is better. 48, 56, 80, 81

Average Precision (AveP) For a binary classification problem, precision is the ability of the model not to label as positive a sample that is negative, and recall is the ability of the model to find all the positive samples. Average Precision (AveP) is the average value of prediction over the intervall of all possible recall values: it is the area under the precision-recall curve. AveP is between 0 and 1, an higher value is better. 48, 56, 80, 81

Committee for Medicinal Products for Human Use (CHMP) EMA's Committee responsible for human medicines, which plays a vital role in the authorisation of medicines. 44, 45

cross-validation (CV) A technique to assess how a model will perform on unseen data. It involves partitioning a sample of data into complementary subsets, performing the analysis on one subset, and validating the analysis on the other subset. 47, 65, 96

Decision Tree (DT) A popular classification model based on recursive partitioning of training examples. 38, 39, 47, 64–66, 79–83, 86–88, 90–94, 96

European Medicines Agency (EMA) Agency of the European Union in charge of the evaluation and supervision of medicinal products. 41, 44, 45, 83

Logistic Regression (LogReg) A popular classification model that uses the sigmoid logistic function to model binary target variables. iv, v, 38, 39, 42, 44, 47, 56–58, 64, 65, 76, 77, 79, 81–83, 86–88, 90–94, 96

Maximum Common Subgraph (MCS) The largest set of atoms and bonds in common between two molecular structures. 50

Molecular Dynamics (MD) A computer simulation method for the analysis of the physical motion of atoms and molecules. It can be used to study a variety of chemical systems, from small molecules, to macromolecules, to parts of cells. In the present work, it is used to simulate the behaviour of antifreeze peptides, and to observe their ice-pinning mechanism, with consequent ice growth arrest. 2–6, 10

Molecular Interaction Field (MIF) A 3D tensor that describes the spatial variation of the interaction energy between a molecule and a chosen probe. 42, 57

Molecular Operating Environment (MOE) A drug discovery software platform, used in the present work to calculate VolSurf descriptors. 42, 46, 57

Principal Component Analysis (PCA) PCA is the process of decomposing a multivariate data-set in a set of successive orthogonal components that explain a maximum amount of the total variance. The original data-set with N features x_1, x_2, x_3, ..., x_N is converted to a new set of features PC_1, PC_2, PC_3, ..., PC_N. The new set of features are the Principal Components (PCs). The PCs have some advantages over the original features. The PCs are orthogonal to each other. The PCs are sorted by the amount of variance that they explain: PC_1 explains more variance than PC_2, PC_2 more than PC_3, and so on. To explain 100% of the original variance, N PCs are required. But usually, the first few PCs are sufficient to explain an high percentage of the original variance. PCA has applications in Molecular Dynamics. It is often used on the C_α coordinates, so each PC represents a global motion of a protein. 15, 22, 24, 25

Protein Data Bank (PDB) The PDB is a database for 3D structural data of biological molecules. It can be freely accessed on the Internet via the websites of its member organizations. The PDB is overseen by an organization called the Worldwide Protein Data Bank (wwPDB). PDB is also the name of the file format commonly used to store 3D structures of biological molecules (extension `.pdb`). The file format is so popular that it is often used as a generic storage of 3D chemical structures. 2, 11, 29, 32

Protein-Ice Contact Surface (PICS) In the present work, PICS is the main measure of peptide-ice interactions. PICS is a per-residue ΔSASA: the per-residue SASA of the peptide minus the per-residue SASA of the peptide-ice complex. Peptide residues with higher PICS interact more with the ice slab. In the present work, PICS values were calculated at each time-frame of the trajectories. The PICS were then averaged over the trajectories, and confidence intervals were calculated with bootstrapping. iv, 15, 18, 19

Quantitative Structure-Activity Relationship (QSAR) The development of models that relate structural properties of molecules (descriptors) to their physicochemical and biological activities. When the activity that is modeled is a physicochemical property (and not a biological activity), QSAR is usually called Quantitative Structure-Property Relationship (QSPR) instead. The term QSAR is more general, and it is used in the present work. 37, 38

Random Forest (RF) A popular ensamble learning model for classification. A Random Forest combines the predictions of a set of Decision Trees to reduce Decision Trees' habit of overfitting to their training-set. 38, 39, 47, 64–66, 79–83, 86–88, 90–94, 96

Rapid Overlay of Chemical Structures (ROCS) A tool for aligning molecules based on 3D shape similarity and distribution of chemical features. It calculates the 3D similarity measure used in the present work: TanimotoCombo. iv, 34–36, 42, 49, 51, 59–63, 97

Root-Mean-Square Deviation (RMSD) The Root-Mean-Square Deviation (of atomic positions) is the main measure of average distance between atoms in different conformations of a molecule. It is defined as: $\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}$, where N is the number of atoms, and d_i is the distance between the atom i in two conformations. It is often calculated after superposition of the two conformations by rotation and translation (without changing the internal coordinates of the conformations). It is popularly used to monitor the evolution of proteins in molecular dynamics simulations, but it can also be used to compare conformations of small molecules. 15, 16, 33

Root-Mean-Square Fluctuation (RMSF) The Root-Mean-Square Fluctuation (of atomic positions) is a measure of the fluctuations of the atoms about their average position. For an atom i, the RMSF is: $\mathrm{RMSF}_i = \sqrt{\langle(\mathbf{x}_i - \langle\mathbf{x}_i\rangle)^2\rangle}$, where $\langle\rangle$ is the time average. RMSF is usually calculated on the C_α atoms of proteins, so RMSF_i can be interpreted as the fluctuation of residue i. Before calculating RMSF, the trajectory structures are usually aligned on a reference frame, or on the average structure. v, vi, 15, 16, 18, 103, 106, 109

Simplified Molecular-Input Line-Entry System (SMILES) A specification for describing the structure of chemical species using strings of characters. A single string of character represents a molecular graph. For instance, the string

"`Cn1c(=O)c2c(ncn2C)n(C)c1=O`" is the SMILES for the caffeine molecule. SMILES are a compact molecular representation: they are very popular in chemical databases, and as input for cheminformatics tools. 30, 45, 113

Solvent-Accessible Surface Area (SASA) The surface area of a molecule that is accessible to a solvent. SASA is typically used to analyze protein structures, and their interactions with ligands. Visually, the SASA is the area that is obtain by rolling a water molecule size sphere above the van der Waals spheres of the molecule in question. This is the total SASA for a molecule. For proteins, the contributions of each residue to the total SASA are typically considered; only externally exposed residues contribute to the SASA, residues hidden in the protein folds do not contribute. It is possible to use SASA as an estimate of the affinity of a protein for a ligand, and to assess what protein residues contribute to the interaction with the ligand. The per-residue SASA of the protein without the ligand, and the per-residue SASA of the protein-ligand complex, are calculated, and then subtracted: this measure is called ΔSASA. Only protein residues that interact with the ligand will have a positive ΔSASA, all the other residues will have zero ΔSASA. 15

Transferrable Intermolecular Potential with 4 Points for Ice (TIP4P/Ice) There are many models that describe the properties of water. Usually, the water models are combined with other force fields, to represent molecular systems in water solutions (e.g., a protein in water). The most popular water models are part of the Transferrable Intermolecular Potential (TIP) family. TIP models describe a water molecule with three points (TIP3P), four points (TIP4P)… Using more points to describe water molecules usually produces a more accurate and realistic representation of the molecular system. In the present work, we have used TIP4P/Ice: a water model based on TIP4P, but with new parameters that better describe the properties of liquid water at cold temperatures, and the properties of water in the solid state (ice). 6, 11, 14, 20

Winter Flounder Antifreeze Protein (wfAFP) The Antifreeze Protein (AFP) from Winter Flounder (an artic fish). It is the most studied AFP of the Type I family. The efficacy of Type I AFPs is moderate when compared to that of other AFPs, but they are appealing for practical applications because of their simple structure (long helices) and relatively short sequences. In the present work, we have modeled peptides derived from the original crystal structure of wfAFP (PDB id 1WFA). 7–11, 16, 18, 20, 25

Glossary

ChEMBL A manually curated database of bioactive molecules with drug-like properties. It is maintained by the European Bioinformatics Institute, of the European Molecular Biology Laboratory (EMBL), based at the Wellcome Trust Genome Campus, Hinxton, UK. In the present work, it was used to retrieve the molecules that compose the new data-set. 29, 48, 49, 95

cross-test In the present work, to cross-test is to use the new data-set to evaluate models built on the original training-set by Franco *et al.*, and vice versa to use the original training-set to evaluate models built on the new data-set. 45, 80, 82, 83, 90, 91, 93, 95

double-feature In the present work, double-feature models are similarity-prediction models built using two input features, specifically Tanimoto CDK Extended and TanimotoCombo. These models combine the predictive power of a 2D and a 3D feature. vii, viii, 47, 53, 64, 65, 73, 76, 79–83, 86–88, 90–94

fingerprint Molecular fingerprints are binary vectors where each bit indicates the presence (1) or absence (0) of a particular substructural fragment within a molecule plural. 33, 34, 42, 44–47, 59, 95

force field In the context of computational chemistry, a force field is a method used to estimate the forces between. It is a functional form and a set of parameters used to calculate the potential energy of a molecular system. Forces described by a force field are usually divided in intramolecular forces (bonds, angles...) and intermolecular forces (electrostatic potential, van der Waals forces...). Other computational chemistry methods depend on force fields, such as conformer optimization, molecular docking, and molecular dynamics. 3–6, 11

hyperparameter An hyperparameter is a parameter whose value is used to control the learning process of a model (it is not automatically derived during the training process as the other parameters). 47, 64–66, 73, 80, 82, 83, 86, 87, 90, 91, 96

NGLview A viewer of 3D molecular structures and trajectories for the Jupyter environment. 51, 97

OMEGA A popular conformer generation tool. It is based on a set of rules to generate conformers similar to solid-state structures of druglike molecules. 32, 33, 46, 97

overfitting Overfitting is to build a model that too closely corresponds to the training-set (usually because the model has learnt the noise in the training-set), and therefore fails to predict external observations reliably. 39, 47, 64–66, 79–82, 86, 90, 91, 96

pChEMBL A standardized measure of bioactivity of the ChEMBL database, suitable for comparing database entries and performing outlier detection. 48, 49

RDKit An open source toolkit for cheminformatics. In the present work, it is used as the main tool to process SMILES strings. 45, 48, 50

regularization Any modification of the learning method of a model to improve performance on an unseen data-set by reducing overfitting. For Logistic Regression, regularization is often achieved by applying a penalty to the loss function during the learning process, thus shrinking the coefficients of the model. For Logistic Regression, two types of penalty (and two types of regularization) are often applied: L1 and L2. L1 regularization uses a $\sum |\beta_j|$ penalty, and L2 regularization uses a $\sum \beta_j^2$ penalty, where β_j are the Logistic Regression coefficients. 47, 64, 65, 79, 81

similarity-prediction In the present work, similarity-prediction is the computational prediction of human assessments of molecular similarity. 38, 42, 44–48, 56, 57, 59, 60, 64–66, 73, 75, 76, 78, 79, 84, 86, 87, 90, 93, 95, 96

single-feature In the present work, single-feature models are similarity-prediction models built using only one input feature. iv, v, vii, 47, 53, 56–59, 64, 65, 73, 76–79, 81–83, 85–94

Tanimoto Tanimoto coefficients are a popular molecular similarity measure, typically calculated on 2D fingerprints. Different fingerprints encode different structural features of a molecule. Tanimoto coefficients calculated on a certain type of fingerprint measure the similarity relative to the structural features encoded by the fingerprint. Tanimoto coefficients are between 0 and 1, with higher values for more similar molecules.. 33, 34, 36, 42, 44–47

Tanimoto CDK Extended In the present work, it is the Tanimoto coefficient calculated on the CDK Extended fingerprint, which is considered the best 2D fingerprint for the similarity-prediction task. iv, v, 47, 49, 50, 56, 57, 59, 60, 64–67, 73, 75–79, 81–89, 93, 113

TanimotoCombo The main 3D similarity measure used by ROCS. TanimotoCombo considers molecular shape and position of chemical features in 3D space. TanimotoCombo values are between 0 and 2, with higher values for more similar molecules. iv, v, 46, 47, 49, 50, 56–60, 64, 66, 67, 73, 75–79, 81–89, 92, 94, 95, 113

test-set A data-set (kept separated from the training-set) used to evaluate the performance of a machine learning model on unseen data. 44, 45, 47, 80

training-set A data-set used to build a machine learning model; the model is "trained" to learn patterns in the data-set. v, vii, viii, 37, 38, 44–47, 50, 53, 56, 57, 59, 64–66, 73, 75–96

VolSurf Molecular descriptors that encode 3D physico-chemical properties. They are calculated on Molecular Interaction Fields using MOE software. iv, 42, 46, 56–58, 78

Symbols

dis2D,dis3D In the present work, this label is assigned to molecule-pairs that are classified as dissimilar by both the 2D and 3D protocols (see subsection 8.4.1). 49, 50, 66, 69, 72, 73, 75, 78, 88

dis2D,sim3D In the present work, this label is assigned to molecule-pairs that are classified as dissimilar by the 2D protocol, and as similar by the 3D protocol (see subsection 8.4.1). 49, 50, 66, 69, 72, 73, 78, 79, 88

sim2D,dis3D In the present work, this label is assigned to molecule-pairs that are classified as similar by the 2D protocol, and as dissimilar by the 3D protocol (see subsection 8.4.1). 49, 50, 66, 69, 72, 73, 75, 78, 79, 88

sim2D,sim3D In the present work, this label is assigned to molecule-pairs that are classified as similar by both the 2D and 3D protocols (see subsection 8.4.1). 49, 50, 66, 69, 72, 73, 75, 78, 88

C_α In the context of proteins and amino acids, the α-carbon (C_α) atoms are the backbone carbone atoms befor the carbonyl atoms. The C_α atoms carry the chemical groups that differentiate the amino acids in proteins. C_α atoms are very important when analyzing the evolutions of protein structures: they approximate the location of each amino acid. iv, x, xi, 12, 15, 16, 22–25

K_i Inhibition Constant (also called Inhibitor Constant). It is the equilibrium constant for the complexation of a reversible inhibitor with its target enzyme. It is commonly used to rank the binding affinities of different molecules for an enzyme. It is a measure of molecular bioactivity. 48

L_{log} The log loss (also called logistic loss or cross-entropy loss), is the score commonly used as the objective function to optimize Logistic Regression models. It can be also used as a performance metric. It basically calculates the difference between ground truth and predicted probability for every observation, and average those errors over all observations. Lower L_{log} values are better, the perfect score is 0. 48, 56, 65, 77, 79–81, 83, 85, 87, 88, 90–92, 94

L_{Brier} The Brier score (L_{Brier}) is a measure of how far the predictions of a classification model lie from the true values. It is basically a mean square error in the probability space. Lower L_{Brier} values are better, the perfect score is 0. 48, 56, 65, 77, 79–81, 83, 85, 87, 88, 90–92, 94

N_{correct} Number of samples that were correctly classified by a model, across all classes. 47, 48, 56, 65, 77, 79–83, 85–88, 90–94

Part I

Antifreeze Peptides

Chapter 1

Theoretical background

1.1 Molecular Dynamics

Molecular Dynamics (MD) is a simulation method for the analysis of the physical motions of atoms and molecules [1]–[7]. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic evolution of the system. The first-ever MD simulation is described in a May 1995 report of the Los Alamos Scientific Laboratory, by Enrico Fermi, John Pasta, and Stanislaw Ulam, with the collaboration of Mary Tsingou [8], [9]. A classical MD simulation is the resolution of Newton's Laws of Motion for a set of interacting particles. For a set of particles with starting positions $\mathbf{r}$ and starting velocities $\mathbf{v}$, the force $\mathbf{f}$ acting on the system is calculated as the negative gradient of the potential energy $U(\mathbf{r})$:

$$\mathbf{f}(\mathbf{r}) = -\nabla U(\mathbf{r}) \tag{1.1}$$

At each time frame, Equation 1.1 is calculated for every particle i of the system, in order to obtain the positions r_i and velocities v_i:

$$\begin{aligned} \frac{\mathrm{d}v_i}{\mathrm{d}t} &= \frac{1}{m_i} f_i \\ \frac{\mathrm{d}r_i}{\mathrm{d}t} &= v_i \end{aligned} \tag{1.2}$$

Where f_i is the force applied to the i-th particle with mass m_i. The same principles can be applied to any system of interacting particles, but when doing MD, the particles usually represent atoms, and the system describes one or more molecules. Classical MD is suitable to study large systems, and it is typically used to simulate the motions of biomolecules such as proteins or DNA [10]–[13]. The starting positions $r_i(0)$ are usually obtained from structural files (see subsection 6.2.2), such as the files downloaded from the Protein Data Bank (PDB). For disordered systems (such as water molecules), the positions can be generated randomly, or an ordered structure can be created and then simulated until it is realistically disordered. The starting velocities $v_i(0)$ are assigned to the atoms according to a Boltzmann distribution. The atomic forces f_i depend on the choice of the force field (see below).

MD simulations are mainly used to predict equilibrium properties of the system, for instance thermal properties, or the available conformations and their relevance in the conformational ensemble. MD is also used to investigate the dynamics of phenomena that could not be easily observed with experimental techniques, for instance the kinetical mechanism of the enzymes.

1.2 Integration of the equations of motion

The molecular systems of interest typically consist of a vast number of particles, and it is impossible to determine the properties of such complex systems analytically; MD simulation circumvents this problem by using numerical methods. However, long MD simulations are mathematically ill-conditioned, generating cumulative errors in numerical integration that can be minimized with proper selection of algorithms and parameters, but not eliminated entirely.
The simplest integrator is the Euler algorithm. It is of very straightforward implementation, but it is usually regarded as very inaccurate. It introduces large errors if the time step used for integration is not much smaller than the smallest intrinsic time scale of the system [7].

In the present work we opted for the "leap frog" algorithm as our integrator of choice, since it is a very popular option among the algorithms usually employed in MD simulations, and the default for Gromacs [14]. The leap frog algorithm uses positions $\mathbf{r}$ at time t and velocities $\mathbf{v}$ at time $t-\frac{1}{2}\Delta t$. The forces $\mathbf{f}(t)$ are determined by the positions $\mathbf{r}(t)$, using the following relations:

$$\begin{aligned}\mathbf{v}(t+\frac{1}{2}\Delta t) &= \mathbf{v}(t-\frac{1}{2}\Delta t) + \frac{\Delta t}{\mathbf{m}}\mathbf{f}(t) \\ \mathbf{r}(t+\Delta t) &= \mathbf{r}(t) + \Delta t\mathbf{v}(t+\frac{1}{2}\Delta t)\end{aligned} \tag{1.3}$$

The algorithm has third order precision in $\mathbf{r}$ and is time-reversible. It can be easily be modified for the inclusion of constraints. It is considered the most suitable for general purposes MD, when very high accuracy is not required.

The typical integration step Δt is 1 fs if the bond lenghts are free to change in the course of the simulation. For normal simulations of big systems, the bond lenghts are constrained, and the Δt can go up to 2 fs, thus halving the number of simulation steps (and the computational time), at the expense of a small loss in accuracy.

1.3 The force field

A simulation is always an approximation of reality. Even if a good integration algorithm and an appropriate Δt are chosen, the errors deriving from numerical integration are inevitable. Beyond the integration errors, the neglection of quantum properties is the physical inaccuracy that comes from the main approximation underlying the typical MD methods. In classical MD, the set of particles is described as a classical system, thus neglecting the quantum properties. Each particle will be described by its three spatial coordinates and by three velocity coordinates, instead that by a field, defined in all points of space. Furthermore, the MD methods more used in the study of macromolecules just consider the atomic nuclei, the individual electrons are neglected. Neglecting the quantum properties means that some interesting phenomena cannot be studied by the classical approximation. For example, chemical reactions cannot be studied, since chemical reactivity comes from the electrons.

The quantum properties being neglected, it becomes necessary to recapitulate the six classical coordinates (r_x, r_y, r_z, v_x, v_y, v_z) in an effective way, by introducing an "effective potential", usually called force field: a mean to implicitly include the main quantum effects when calculating the forces acting on the particles using just the classical degrees of freedom. There is arbitrariness in this: many force fields are available in the literature [15]–[17]. Usually, each force field is developed whit a particular class of molecules in mind. Many force fields are optimized for use on biological molecules.

A force field is a mathematical expression describing the dependence of the energy of a system on the coordinates of its particles. A force field consists of an analytical form of the interatomic potential energy (the functional form) $U(\mathbf{r})$, and a set of parameters entering into this form. The parameters are typically obtained either from ab initio or semi-empirical quantum mechanical calculations or by fitting to experimental data such as neutron, X-ray and electron diffraction, NMR, infrared, Raman and neutron spectroscopy. Molecules are defined as a set of atoms held together by simple harmonic forces. Ideally a functional form should be simple enough to be computed efficiently, but sufficiently detailed to reproduce the properties of interest of the modeled system.

1.3.1 CHARMM27

Force fields differ in the degrees of complexity (the number and kinds of terms in the functional form), the methods used to obtain the parameters, and the kinds of systems that can be modeled. In the present work, we used CHARMM27 force field [18], because it was the most appropriate to model the system that we wanted to examine (see section 3.1). The functional form of CHARMM27 is reported in Equation 1.4:

$$
\begin{aligned}
U(\mathbf{r}) = & \sum_{\text{bonds}} K_b(b-b_0)^2 \\
& + \sum_{\text{UB}} K_S(S-S_0)^2 \\
& + \sum_{\text{angles}} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{\text{dihedrals}} K_\chi(1+cos(n\chi-\delta)) \\
& + \sum_{\text{impropers}} K_\varphi(\varphi-\varphi_0)^2 \\
& + \sum_{\text{non-bonded}} \left\{ \epsilon_{ij}\left[\left(\frac{R_{ij}^{\text{min}}}{d_{ij}}\right)^{12} - \left(\frac{R_{ij}^{\text{min}}}{d_{ij}}\right)^{6}\right] + \frac{q_i q_j}{e d_{ij}} \right\}
\end{aligned}
\tag{1.4}
$$

As most force fields, the CHARMM27 includes bonded and non-bonded terms. Bonded terms describe the intramolecular forces, whereas non-bonded terms describe the intermolecular forces. The intramolecular terms depend on the bond lenght b, the distance between two covalently bonded atoms (1-3 distance) S, the valence angle θ, the dihedral (torsion) angle χ, and the improper angle φ. The forces for bond lenght, 1-3 distance, valence angle, and improper angle, depend on two parameters each: a force constant K and an equilibrium value (b_0, S_0, θ_0, and φ_0). The dihedral force depends on three parameters: the force constant K_χ, the multiplicity n, and the phase angle δ.

CHARMM27 uses the most common terms for the intermolecular non-bonded interactions: the Coulomb term for electrostatic forces, and the Lennard-Jones term for the van der Waals forces. Both terms depend on the distance between two non-bonded atoms, d_{ij}. The Coulomb term has one parameter (for each atom): the atomic charge q_i. The Lennard-Jones term has two parameters (for each pair of atoms): the well depth ϵ_{ij} and the minimum interaction radius R_{ij}^{min}.

CHARMM27 is an advanced force field, developed after years of experience utilizing previous force fields. The direct predecessor of CHARMM27 is CHARMM22. CHARMM22 had some limitations, most importantly it overstabilized the A form of DNA. In order to overcome the limitations, CHARMM27 was built with better parameters. The new parameters are obtained with a more complex optimization procedure, that take into account the intrinsic energetic properties of a set of model compounds, and the overall conformational properties of DNA and RNA. The target dataset is mixed: it includes computed Quantum Mechanical properties, and experimental condensed phase properties. The new parametes work well in a variety of environments, and CHARMM27 is a good general purpose force field for MD simulations of biomolecules.

1.3.2 TIP4P/Ice water model

The majority of force fields for biomolecules (including CHARMM27) only have parameters to describe proteins, DNA, RNA, and lipids. The force fields themselves do not have parameters for water molecules. In reality, biomolecules are in aqueous environments. To describe biomolecular systems realistically, it is necessary to simulate them in the presence of water. Many water models are available in the literature (and in the most popular MD software tools) [19]–[21]. Water models are generally independent on the choice of the biomolecular force field. CHARMM27 was optimized to be used with the very popular TIP3P water model, but it can be used successfully also with other water models.

The most popular water models are part of the Transferrable Intermolecular Potential (TIP) family. TIP models describe a water molecule with three points (TIP3P), four points (TIP4P)... Using more points to describe water molecules usually produces a more accurate and realistic representation of the molecular system. In the present work, we have used TIP4P/Ice: a water model based on TIP4P, but with new parameters that better describe the properties of liquid water at cold temperatures, and the properties of water in the solid state (ice) .

All the TIP models have the same basic geometry: they have a d_{OH} distance of 0.9572 A, and a θ_{HOH} angle of 104.52°. The TIP models are rigid: d_{OH} and θ_{HOH} do not change in the course of the simulations. All TIP models have Lennard-Jones parameters for the oxygen atom, and a positive atomic charge (the Coulomb parameter) $q_{\mathrm{H}}$for the hydrogen atoms. The specific values for the Lennard-Jones and Coulomb parameters vary among the TIP models. Since a water molecule is neutral, all TIP models also have a negative charge, equal to $-2q_{\mathrm{H}}$, to counterbalance the hydrogen charges. For TIP3P (the simplest TIP model), the negative charge is located on the oxygen atom. For TIP4P, the negative charge is located on a point M, located at a distance d_{OM} along the θ_{HOH} bisector. The distance d_{OM} is another parameter of the water model. The TIP4P parameters are optimized to reproduce the vaporization enthalpy and density of liquid water at room temperature. For this reason, TIP4P is suitable to simulate aqueous environments at room temperature, and it is a very popular water model.

Since TIP4P was optimized to reproduce the properties of liquid water at room temperature, it is not a very realistic model for water at cold temperatures, close to the ice melting point. To simulate water at cold temperatures, a variation of the TIP4P model was developed: the TIP4P/Ice model [22]. Instead of only reproducing water properties at room temperature, TIP4P/Ice parameters were obtained to also fit the melting lines and coexistence lines of different ice forms. When used in simulations, TIP4P/Ice gives a better phase diagram and better densities of several ice forms than the other TIP models. TIP4P/Ice is suitable to simulate icy environments.

1.4 Antifreeze proteins

The production of Antifreeze Proteins (AFPs) is an important strategy organisms have developed to thrive in cold ecosystems where there is a risk of freezing [23], [24]. The typical mechanism of action of AFPs is to bind to ice surfaces to control the spontaneous ice growth that would occur without AFPs. For this reason, AFPs that bind to ice surfaces are often called Ice-Binding Proteins (IBPs).

The structure of AFPs found in different organisms is remarkably diverse, considering that AFPs share the same ligand [25]. This diversity reflects the independent origin of AFPs on numerous occasions during the course of evolution. The crucial activity of AFPs is thermal hysteresis, the depression of the freezing point of water below the melting point. Some AFPs have relatively weak thermal hysteresis activity, and help organisms tolerate freezing through other processes, such as ice recrystallization inhibition.

The AFPs with the simplest structure are the so called Type I AFPs [26]. Type I AFPs are alanine-rich (> 60%), possess an high helical content, and contain 11-residue repeat sequences that start with threonine. The most well-studied protein in this group is the liver-isoform from Winter Flounder (gene: HPLC6). This Winter Flounder Antifreeze Protein (wfAFP) has 37 residues. It contains three 11-amino acid repeats of the sequence TX2NX7, where X is usually an alanine or another amino-acid that favours α-helix formation. Conformational studies have shown that the protein is completely α-helical, with the exception of the last unit which adopts a 3_{10}-helix conformation [27]. The N-terminal part of the protein is an elaborate cap structure, which is likely to contribute to the stability of the helix. This N-terminal cap consists of an ordered network of eight hydrogen bonds involving the side chains of ASP1, THR2, SER4, and ASP5 residues.

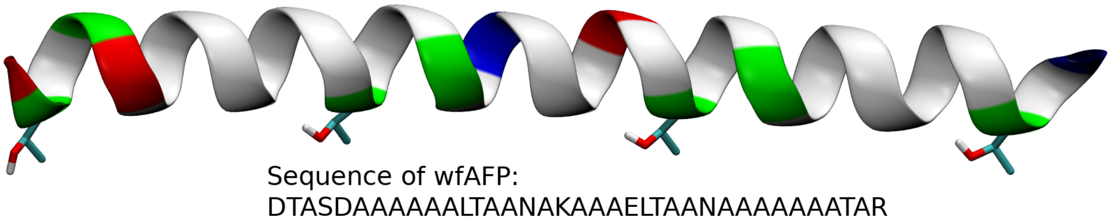
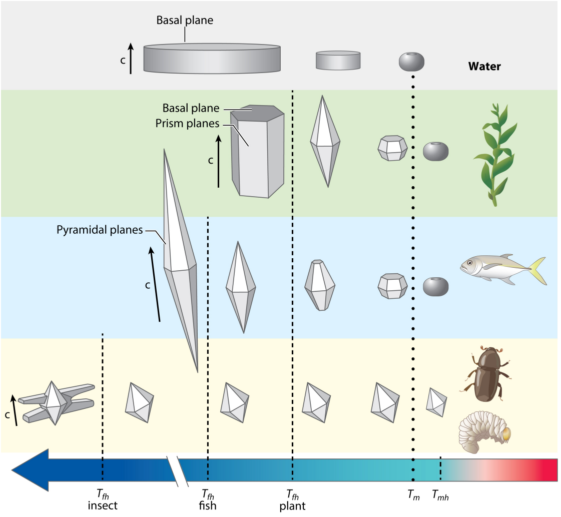


Figure 1.1: Crystal structure and residue sequence of Winter Flounder AFP (gene: HPLC6). Backbone colored by residue type, THR sidechains in evidence.

The different types of AFPs bind to different interfaces of ice crystals. (Figure 1.2). The wfAFP binds to the $\langle 01\bar{1}2 \rangle$ axis on the pyramidal plane of the I_h ice form (the ordinary ice) [27].

Computational studies have played an important role in explaining the mechanism of action of the wfAFP. Most of the AFPs with an helical structure have two sides: a



Bar Dolev M, et al. 2016.
Annu. Rev. Biochem. 85:515–42

Figure 1.2: Different AFPs bind to different ice crystal planes, and present different levels of thermal hysteresis activity [24].

polar side, and an apolar one. It was originally thought that the polar side of wfAFP interacted with ice. But computational studies have shown that the contrary is true [28]. The hydrophobic side of the wfAFP can approach ice more closely, thus participating in kinetic pinning leading to its antifreeze activity. On the other hand, the polar side of the protein interacts with liquid water, thus helping to keep it in its liquid form.

1.5 Note

The following chapters are taken from:
E. Gandini, M. Sironi, and S. Pieraccini, "Modelling of short synthetic antifreeze peptides: Insights into ice-pinning mechanism," *Journal of Molecular Graphics and Modelling*, vol. 100, p. 107 680, Nov. 2020, ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2020.107680

Chapter 2

Introduction

Antifreeze Protein (AFP) are a class of structurally diverse proteins that protect different species of living organisms from fatally freezing in icy environments [24], [30]. AFPs depress the freezing point of water in a kinetic, non-colligative manner, caused by AFP adsorption to specific surface planes of seed ice crystals [31]. Moreover, AFPs inhibit Ostwald ripening recrystallization of ice, preventing the growth of larger crystals and the concomitant shrinking of smaller ones [32]. The action of AFPs is usually rationalized with the adsorption-inhibition model. It assumes that proteins bind irreversibly to ice thus arresting the growth of the crystal at supercooled conditions through the creation of a metastable curved ice surface, according to the Kelvin effect [33]. Because of their unique capability in controlling ice formation, AFPs are very attractive for potential practical applications including food storage, anti-icing coatings for vehicles and infrastructure, and cryopreservation of cells and tissues [34], [35]. Their potential uses has prompted interest in research and led to numerous experimental and computational studies [36]–[43]. However, the molecular complexity, the limited availability and the consequent costs, hamper progress toward AFPs practical application [23], [44], [45].

Many different AFPs have been identified and categorized according to their structure and their binding specificity. Among them, Type I AFPs have been widely studied, particularly the liver-isoform from Winter Flounder (gene: HPLC6). Winter Flounder Antifreeze Protein (wfAFP) is a monomeric, alanine-rich protein, composed of 37 residues, with three 11-residues repeats (TA2NA7). Even though the antifreeze effect brought about by wfAFP is moderate when compared to that of other AFPs, it is appealing for practical applications because of its simple structure and relatively short sequence. wfAFP mechanism of action was thoroughly studied both at the experimental and computational level [26]–[28], [46]–[49]. It was demonstrated that wfAFP binds preferentially along the $\langle 01\bar{1}2\rangle$ axis of the pyramidal ice plane [50], [51]. The hydrophobic face of wfAFP is presented to the pyramidal ice plane, and the interaction mechanism is similar to the hydrophobic solvation effect [28].

Efficient synthetic analogues mimicking the effects of AFPs are highly desirable. Three 12-residue analogues of wfAFP (Table 2.1) have been synthesized and successfully applied for the fabrication of anti-icing surfaces by Zhang *et al.* [35]. The present paper is focused on the application of Molecular Dynamics (MD) [1]–[7] based techniques to investigate at the atomic level the antifreeze activity of the three aforementioned peptidic wfAFP analogues and to model their binding to ice surface and their mechanism of action.

Peptide	Sequence	Molecular Weight / Da	Average Freezing Temperature / °C
1–1	DTASDAAAAAAL	1047	-11.9
1–2	DTASDAKAAAEL	1162	-16.3
1–3	DTASDAFAAAAL	1123	-10.8

Table 2.1: Amino acid sequences of the three antifreeze peptides considered in the present study. Residues that differ from original wfAFP are underlined. The table also includes molecular weights, and reported average freezing temperatures of water droplets on peptide-coated silicon wafers.

Chapter 3

Methods

3.1 Modeling of the peptides

Three 12 residues peptides derived from wfAFP have been considered (Table 2.1). Peptide 1−1 is composed of the first 12 residues of wfAFP. The wfAFP crystal structure was obtained from PDB ID 1WFA [27], [52]. The relevant residues were kept, all the others were manually removed. Peptides 1−2 and 1−3 present residue mutations that were designed to improve their antifreeze properties [7]. In particular, for Peptide 1−2, Alanine 7 was mutated to a Lysine, and Alanine 11 to a Glutamic Acid, in order to form an intra-molecular saline bridge to increase helicity. For Peptide 1−3, Alanine 7 was mutated to a Phenylalanine, in order to improve hydrophobic interactions with ice. The structural models of Peptides 1−2 and 1−3 were obtained through residue mutation with UCSF Chimera Rotamer tool [53], [54].The peptides were then protonated with Gromacs 5.0.4 [14], and described with CHARMM27 force field [18]. The CHARMM27 force field was chosen because it shows the best performance in modeling ice-protein interactions along with the TIP4P/Ice water model [22], [55]. The three antifreeze peptides, and control peptides dodeca-Glycine (G12) and dodeca-Alanine (A12), were then subjected to three kinds of simulations.

3.2 Molecular Dynamics Simulations

3.2.1 Simulations in a water box

The peptides were inserted into a cubic box of side 4.5 nm. The box was solvated with TIP4P/Ice [22] water molecules, and brought to charge neutrality with 1 M concentration of Na and Cl ions. The system was then subjected to restrained relaxation and NVT and NPT equilibration, at temperature 275 K and 1 bar pressure. After the equilibrations, the structure restraints were removed, and the peptides were subjected to

1.5 μs molecular dynamics simulations at NPT conditions.

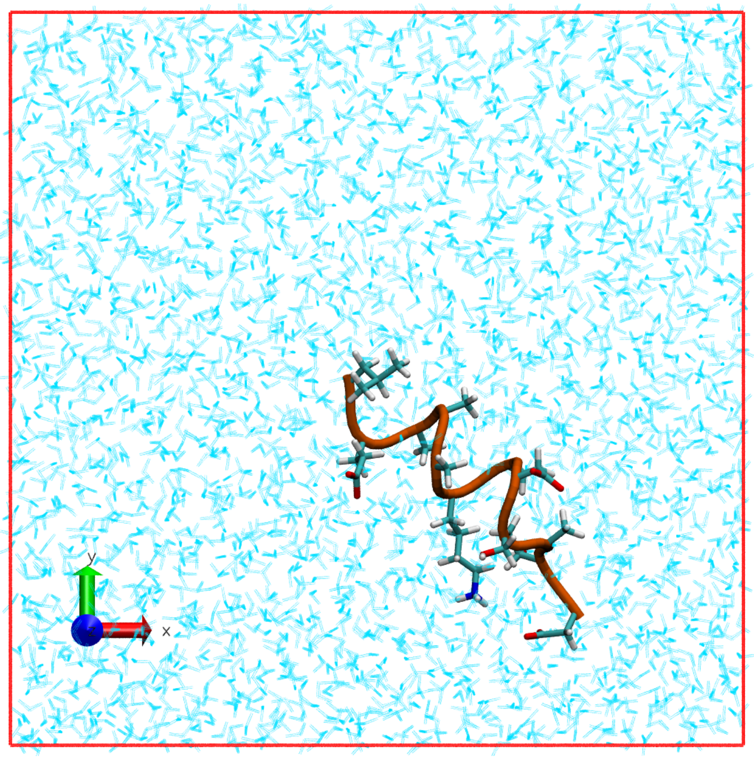


Figure 3.1: Final frame of Peptide 1–2 water box simulation.

3.2.2 Simulations on a fixed-ice surface

A large unit cell of I_h proton-disordered ice was generated with GenIce tool [56]. A 2.5 nm thick slab of ice, exposing the pyramidal plane, was obtained with Vesta [57]. Peptide 1–1 was manually placed on the pyramidal plane along the $\langle 01\bar{1}2\rangle$ axis, with the hydrophobic residues facing ice [28]. Peptides 1–2 and 1–3 were generated in loco through mutation with Chimera. The ice-peptide systems were then inserted in 5.7 × 5 × 16 nm box. Water molecules and ions were then added, but the last 5 nm of the box in the Z direction were left empty, in order to create an ice–peptide–water–vacuum system, as described by Mochizuki *et al.* [58], and shown in Figure 3.3.

Restraints were added to the protein C_α atoms and to the ice oxygens, in order to carry out system relaxations and NVT equilibrations at 275 K. Afterwards, the protein restraints were removed, and the systems were subjected to 100 ns molecular dynamics simulations.

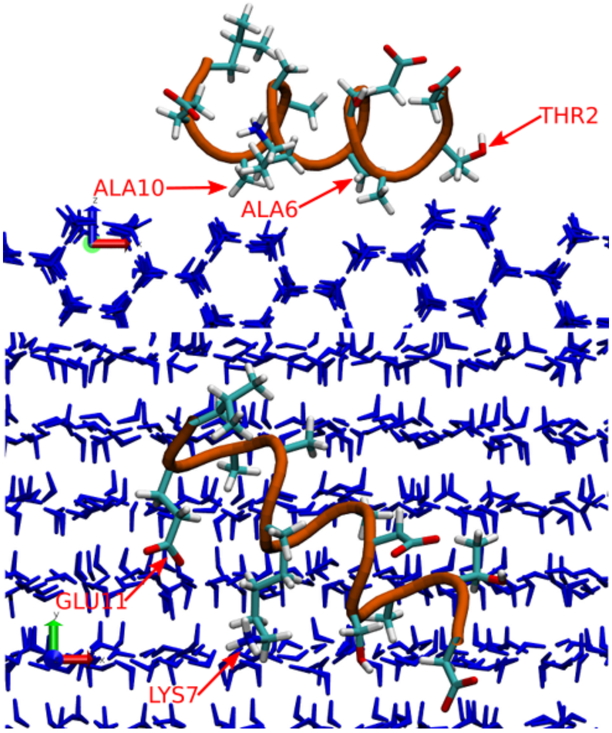


Figure 3.2: Views of Peptide 1–2 fixed-ice simulation final frame. Restrained ice is in blue licorice representation. The backbone is in orange tube representation, and side chains are licorice colored by atom type. For simplicity, liquid water is not shown in the picture (although present in the simulated system).

3.2.3 Simulations on a growing-ice surface

The ice–peptide–water–vacuum systems were prepared in the same way as described above for the fixed ice simulations. In this new set of simulations, the NVT equilibration and the production molecular dynamics simulation were carried out at 248 K, below the reported 270 K freezing point of TIP4P/Ice water model freezing point [22], so that ice growth could be observed within the 850 ns long production trajectories. A simulation without any peptide was included for control.

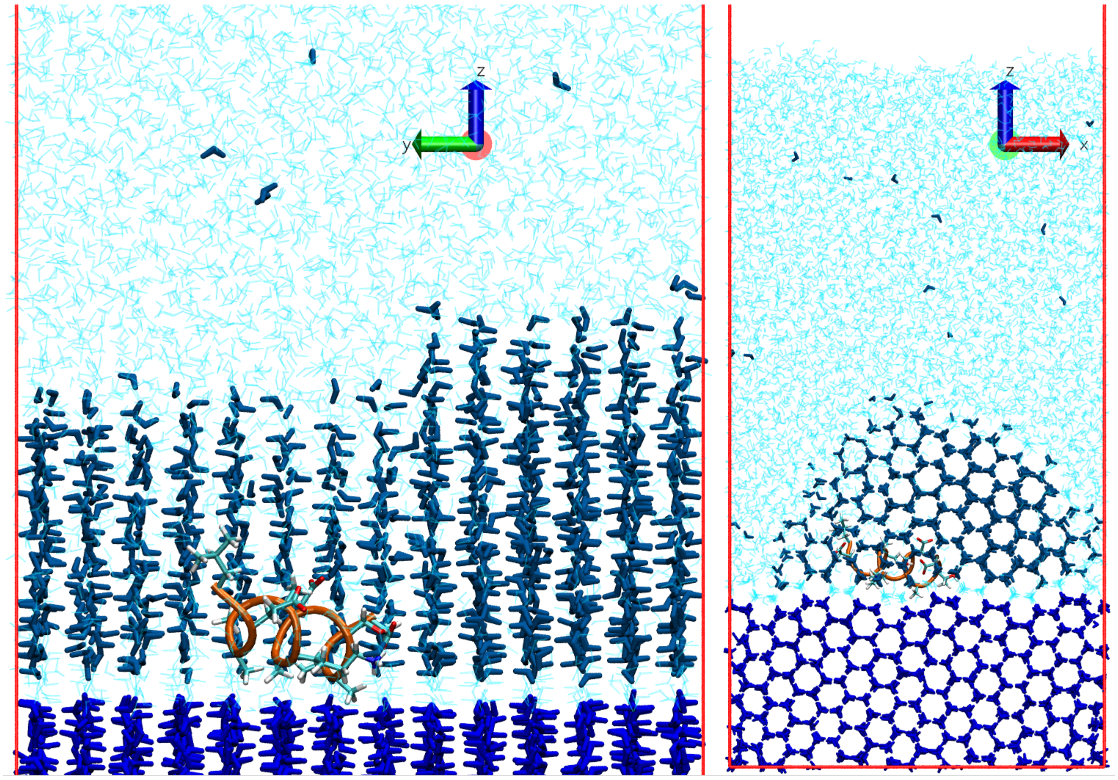


Figure 3.3: Views of the final frame of the growing-ice simulation of Peptide 1–2. Restrained ice is in dark blue licorice representation, whereas ice that was formed from unrestrained water during the simulation is in lighter blue licorice. Liquid water molecules are represented as semi-transparent lines. Protein backbone is in orange tube representation, and side chains are in licorice colored by atom types. Notice the empty space (vacuum) on top of the water molecules on the right view.

3.3 Trajectory analysis

All structural renderings were performed with VMD [59]. Secondary structures were calculated using DSSP [60] algorithm as implemented in MDTraj [61]. Structural RMSD, radius of gyration, helix length and helicity were calculated with Gromacs `rmsd`, `gyrate`, and `helix` tools [14]. Root-Mean-Square Fluctuation (RMSF) calculations were performed with MDAnalysis [62], [63]. Protein-Ice Contact Surface (PICS) were calculated through Solvent-Accessible Surface Area (SASA) with MDAnalysis. Quantification of ice formed during growing ice simulations was performed with CHILL+ algorithm [64]. All graphs were created with Python plotting libraries [65]–[67].

Principal Component Analysis (PCA) [68], [69] was applied to antifreeze and control peptides trajectories in the three sets of simulations. In order to apply PCA, for each simulation set, we obtained the coordinates of C_α atoms throughout each peptide trajectory, using Gromacs `trjconv` tool, and we aligned C_α geometries with MDAnalysis. We then used ENCORE [70] to perform PCA on C_α trajectories of all peptides, for each simulation set. ENCORE concatenates C_α trajectories of all peptides (possible since every peptide has the same number of C_α atoms, twelve in this case), and then applies PCA algorithm implemented in scikit-learn [71] to the concatenated C_α coordinates. Since the C_α atoms trajectories are concatenated in a single coordinate matrix, PCA algorithm calculates the Principal Components (PCs) that describe the total conformational variance across all the peptides. When PCA is applied on concatenated trajectories of different peptides, the calculated PCs highlight conformational differences and similarities between the peptides. Each PC describes a certain percentage of the total conformational variance across all trajectories. The higher this percentage variance, the most important is the global motion described by the PC.

Chapter 4

Results and discussion

4.1 Analysis of water box simulations

The water box set of simulations was tested for conformational stability and helicity of the peptides, which in case of wfAFP is highly correlated with antifreeze activity [24]. To test whether the peptides present the same characteristic helical structure of the original wfAFP, secondary structures were calculated with the DSSP algorithm throughout the trajectories (Figures A.3, A.4, and A.5). All the peptides present high degrees of helicity. Peptide 1–2 is more helical than the other peptides, in good agreement with reported circular dichroism spectra [35]. Moreover, four structural properties of the conformational ensembles of the antifreeze and control peptides (Radius of Gyration, Structural RMSD, RMSD from ideal helix, helix length) were calculated throughout the trajectories and are reported in Figure 4.1. The four chosen properties are measures of conformational stability and resemblance of the peptides with an ideal a helix. In particular, in agreement with experimental results, Peptide 1–2 shows greater propensity for helical structure than Peptides 1–1 and 1–3. C_α atoms RMSF were also calculated on water box trajectories (Figure A.6). Confidence intervals are $2 \times$ Standard Error of the Mean (SEM) calculated with bootstrapping [72], [73]. The residues of the three antifreeze peptides have low RMSF values, indicating limited flexibility. Peptide 1–2 has the lowest RMSFs.

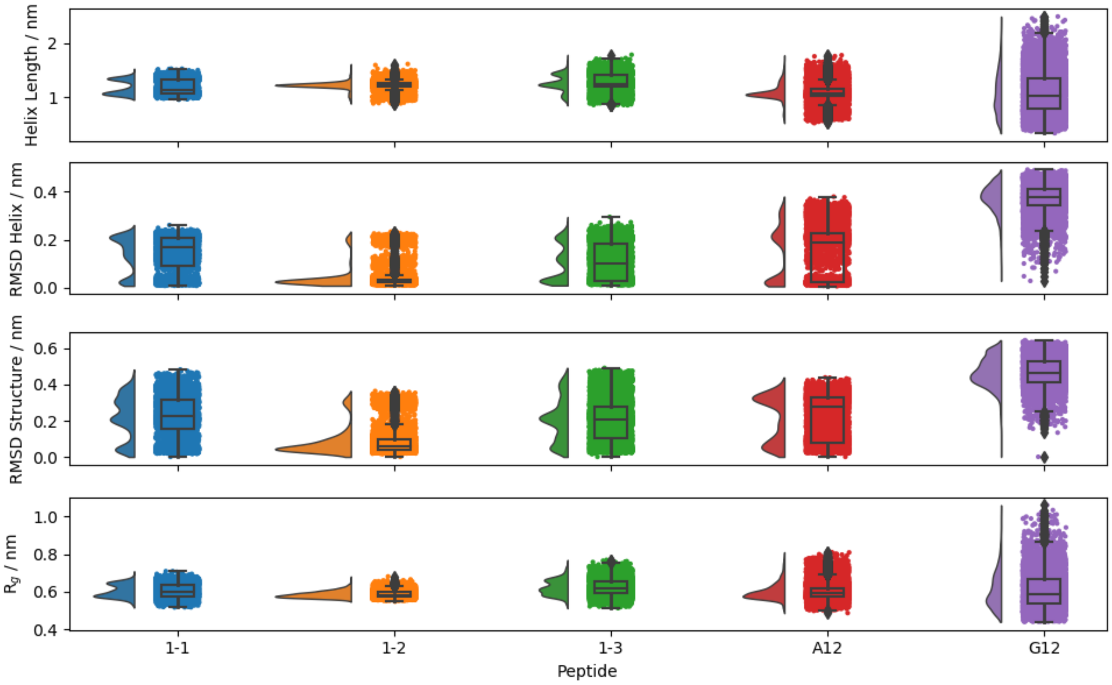


Figure 4.1: Distributions of four structural properties calculated throughout the water box simulations.

4.2 Analysis of fixed-ice simulations

RMSF values were also calculated for the fixed-ice simulations (Figure A.10): they are even lower than in water box simulations, indicating that the presence of the ice surface has a stabilizing effect on peptide conformations. RMSF values does not significantly differ between antifreeze peptides in the fixed-ice simulations. The antifreeze mechanism and the nature of the ice-wfAFP interactions are the subject of many hypotheses [26]. It is believed that both direct interactions between wfAFP with ice surface, and interactions mediated by water molecules at ice-protein interface play a role in the antifreeze activity.

Average PICS was calculated and results are presented in Figure 4.2. Residues with high PICS values are the ones that interact more steadily with ice. THR2, ALA6 and ALA10 interactions are relevant in all three peptides, in good agreement with the observations of Kun and Mastai [74]. Residue 7, which is an ALA for Peptide 1−1, a LYS for Peptide 1−2, and a PHE for Peptide 1−3, also exhibits a large PICS for all the three peptides, but it is significantly larger for Peptide 1−2, suggesting that LYS side chain favors the interaction. Close inspection of the trajectory reveals that the interaction is brought about mostly by carbon atoms of the LYS side chain, and not by its positively charged group, which interacts either with GLU11 or the solvent molecules, as shown in Figure 3.2, which presents two different views of the final trajectory frame of Peptide 1−2 on fixed-ice. The three antifreeze peptides remained firmly attached to the ice surfaces for the duration of the fixed-ice simulations, whereas non-antifreeze Peptide G12, after few nanoseconds, starts to change its secondary structure and to detach from the ice surface. Non-antifreeze Peptide A12 keeps its helical structure throughout the fixed-ice simulation. Structural properties of the five peptides calculated throughout the fixed-ice trajectories are reported in Figure A.9.

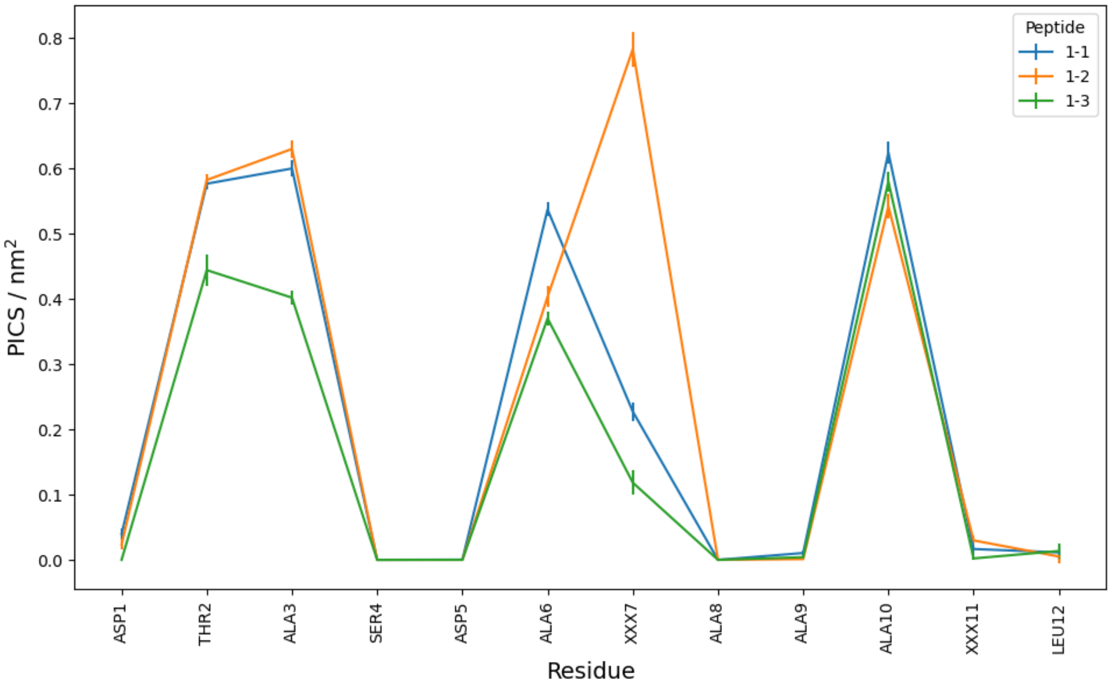


Figure 4.2: Average PICS of antifreeze peptides with bootstrapped error bars, calculated during fixed ice simulations. Residues which are different among the peptides are labeled XXX.

4.3 Analysis of growing-ice simulations

In order to assess the ability of antifreeze peptides to inhibit ice growth, we performed a set of simulations at temperature below the freezing point of TIP4P/Ice. Figure 3.3 presents different views of the final trajectory frame of Peptide 1–2 growing-ice simulation. The peptide is firmly attached to the original restrained ice surface, and it induces the formation of a curved ice front from the liquid water molecules in the course of the simulation, as observed in previous computational studies [40], [75]. No ice growth above the peptide is detected. Quantification of ice formed during growing ice simulations was performed with CHILL+ algorithm [64], and results are shown in Fig. 3(c). Ice starts growing immediately in control simulations performed without peptide: all water is turned into ice after 150 ns. In simulation performed with non-antifreeze peptide G12, ice starts growing more slowly, and the growth is complete after 500 ns. Peptide G12 is completely enclosed in a block of ice formed by initially liquid water. In presence of non-antifreeze Peptide A12, ice growth is slower, and it is complete after 800 ns. Ice is not able to grow above the antifreeze peptides. The total quantity of ice formed with antifreeze peptides is around 10%, throughout the whole 850 ns simulations. Visual inspection of the growing ice trajectories along with the application of CHILL+ algorithm confirmed that the synthetic analogues of wfAFP can shape ice surface inducing the formation of a curved ice front and consequently block ice growth with a mechanism compatible with the Kelvin effect.

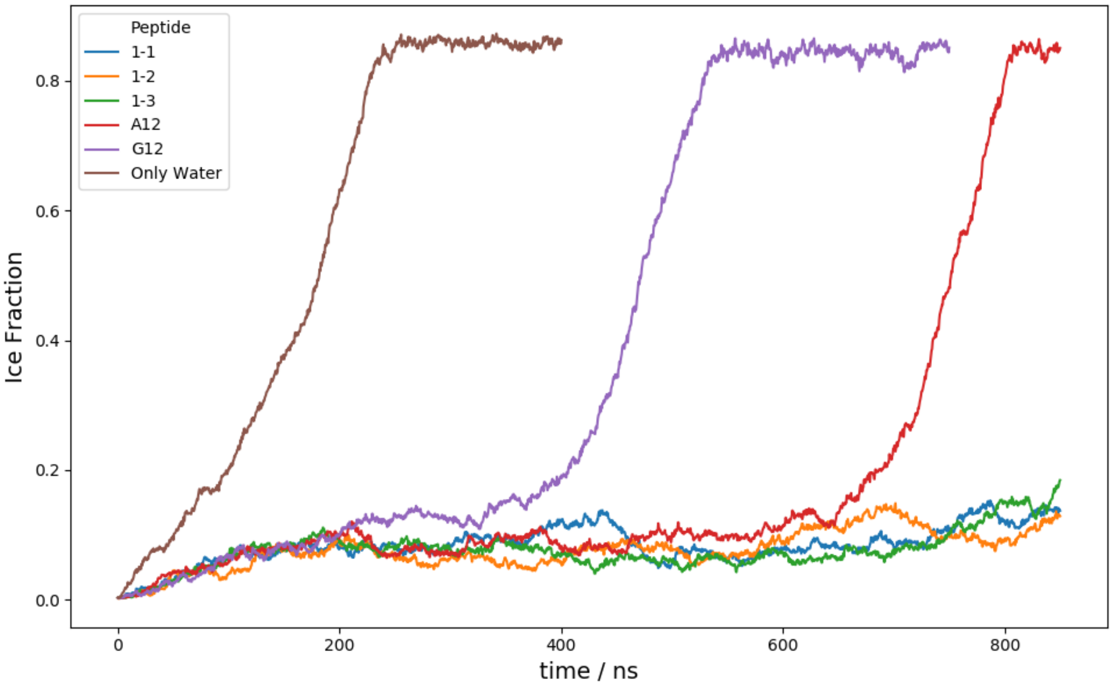


Figure 4.3: Fraction of ice formed from unrestrained water throughout the growing-ice simulations.

4.4 PCA of conformational ensembles

Principal Component Analysis (PCA) is a general purpose statistical procedure [68], [69] that has often been applied successfully to study molecular dynamics trajectories [49], [73], [76]–[78]. When studying a series of similar peptides or proteins, it is interesting to take into account the conformational effects that are induced by residue mutations which can either conserve or modify the conformations of the protein under consideration. In particular, for antifreeze proteins mutation analysis is very important, since interactions with ice strongly depend on protein conformations [79], [80]. PCA is a useful technique to compare a series of same-length similar peptides or proteins. In a single graph, it shows differences in the global motions that are caused by residue mutations, and highlights structural peculiarities of the most effective peptides.

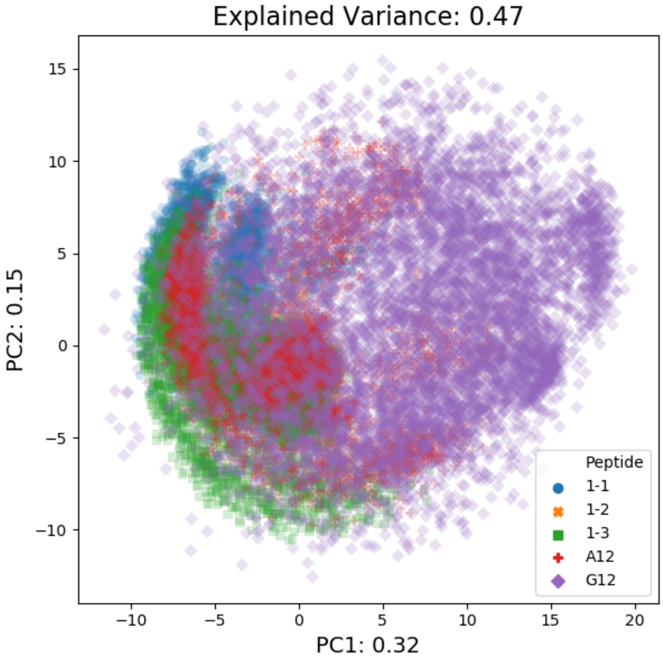


Figure 4.4: C_α coordinates projected onto the first two PCs of water box simulations.

In the present work, conformational ensembles of antifreeze peptides, and of non-antifreeze peptides G12 and A12, were compared through PCA using the algorithm as implemented in ENCORE [47]. The software concatenates C_α atoms coordinates of all peptides throughout the trajectories in a single coordinate matrix, and performs PCA. Figure 4.4 shows C_α atoms coordinates throughout the water box simulations projected onto the first two Principal Components (PCs), that together account for 47% of the total motions. Antifreeze peptides, and non-antifreeze Peptide A12, cover a similar area, thus confirming that they explore a similar conformational ensemble.

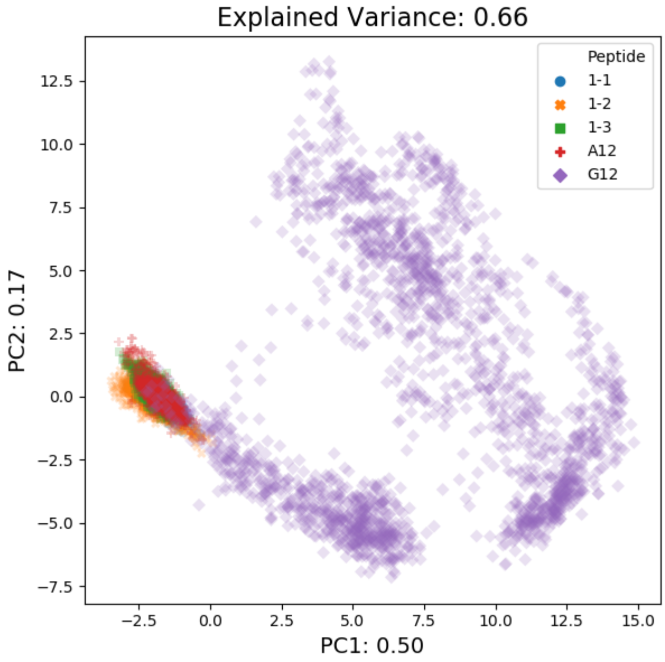


Figure 4.5: C_α coordinates projected onto the first two PCs of fixed-ice simulations.

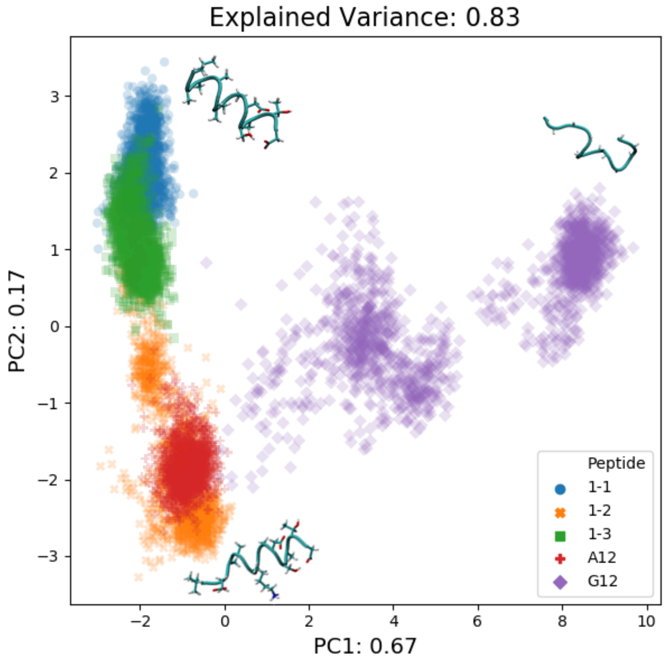


Figure 4.6: C_α coordinates projected onto the first two PCs of growing-ice simulations.

As expected, Peptide G12 covers a much larger area, since it does not have a preferred secondary structure.

PCA was then applied to C_α coordinates throughout the fixed-ice simulations, and results are shown in Figure 4.5. The first two PCs account for 66% of the total motions, which is more than the variance explained by the first two PCs in the water box simulations, because of the stabilizing presence of the ice surface. Antifreeze peptides, and control peptide A12, occupy an even smaller area compared to that covered by G12, pointing out that most of the total variance explained by the two main PCs comes from G12, whereas the antifreeze peptides are much stabilized by the presence of the ice surface. Even though A12 is not an antifreeze peptide, Alanine residues are known to interact with ice [28], [81]. So, A12 is stabilized by the ice surface, and its area is the same size as that of antifreeze peptides.

Results of the application of PCA to growing-ice simulations are shown in Figure 4.6. The variance explained by PC_1 is 67%, whereas that explained by PC_2 is 17%. The total motions explained by the first two PCs is 83%, even higher than the fixed ice simulation, since now the ice growth further limits the conformational freedom of Peptide G12. PC_1 is responsible for the separation of G12 from antifreeze peptides, whereas PC_2 is able to differentiate between Peptide 1–2 and peptides 1–1 and 1–3. Conformations representative of the extreme PC values were extracted from the trajectories and reported onto the PC graph. PC_1 separates a helical structure from a disordered structure. Interestingly, the much smaller global motion represented by PC_2 separates a perfect helical structure from a helical structure with the C-term residues outstretched. A close inspection of Peptide 1–2 structure with outstretched C-term residues, suggests that the elongation may be brought about by GLU11 interaction with liquid water. LEU12, though outstretched, is still interacting with the ice surface, as well as the aliphatic carbons of LYS7, whereas the positively charged group of LYS7 side chain interacts with liquid water. Non-antifreeze Peptide A12 has PC_1 values similar to those of antifreeze peptides, thus confirming that antifreeze peptides and A12 explore a similar conformational ensemble, and have similar secondary structures. On the other hand, Peptide A12 has PC_2 values similar to those of Peptide 1–2.

Chapter 5

Conclusion

Three synthetic analogues of wfAFP, which have shown experimentally measurable ice growth inhibition activity, were the subject of computational modeling and simulation. Three simulation setups were devised, in order to analyze different molecular properties that affect antifreeze activity. Simulations in a water box were able to reproduce the experimentally observed conformational and secondary structure stability of the three antifreeze peptides. Simulations on a fixed-ice surface pointed out the presence of stabilizing interactions between the antifreeze peptides and an ice slab exposing the pyramidal plane. Simulations on a growing-ice surface were able to reveal an ice-growth blocking effect for the three antifreeze peptides. PCA of C_α atoms coordinates pointed out differences in the global motions of antifreeze peptides from non-antifreeze Peptide G12. Secondary structure of Peptide A12 is similar to that of anti-freeze peptides, and PCA confirmed that global motions of A12 are similar to those of antifreeze peptides. Anyway, helical propensity is not, in itself, a guarantee of antifreeze activity [79]. CHILL+ measurements of ice formed during simulations confirm that A12 is unable to block ice growth, even though it is structurally similar to antifreeze peptides. When applied to the growing ice simulation, PCA was also able to extract a structural pattern peculiar to the conformational ensemble of antifreeze Peptide 1–2 that will be useful to design new synthetic analogues of wfAFP.

The protocol that was described in this work is useful to analyze the conformational properties and antifreeze activity of series of short peptides. Even though peptides derived from wfAFP may not present the greatest antifreeze activity at low concentrations, their short chains and simple structures make them promising for large-scale synthesis and practical applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in the previous chapters.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support (application code: HP10C8VXWW). We thank Prof. Sara Pellegrino, Prof. Matteo Pecchiari, Prof. Fabio Acocella, and Dr. Giuliano Zanchetta, for helpful discussion.

Part II

Molecular Similarity

Chapter 6

Theoretical background

6.1 A brief introduction to cheminformatics

Cheminformatics (also popularly spelled chemoinformatics) is very broadly defined as the application of information technology to chemistry. Cheminformatics encompass the design, creation, organisation, storage, management, retrieval, analysis, dissemination, visualisation, and use of chemical information [82]. The field of cheminformatics include many well established techniques, the result of decades of scientific research [83], [84]. The methodologies and infrastructures most commonly used in cheminformatics have been reported by Bunin *et al.* [85], and are listed below:

- Chemical data collection, analysis, and management.
- Data representation and communication.
- Database design and organization.
- Chemical structure and property prediction (including drug-likeness).
- Molecular similarity and diversity analysis.
- Compound or library design and optimization.
- Database mining.
- Compound classification and selection.
- Qualitative and quantitative structure-activity or structure-property relationships.
- Information theory applied to chemical problems.

- Statistical models and descriptors in chemistry.
- Prediction of *in vivo* compound characteristics.

In the present work, we will be focusing on molecular similarity analysis, and on the creation of related statistical models. We will also manage the generation of the dataset needed to build the statistical models. This chapter is not meant to be a thorough exposition of the fundamentals of cheminformatics. For that, there are excellent textbooks [84]–[87]. For a quick overview of the field of cheminformatics, I suggest the one by Engel [88]. In the next sections, we will only present the theoretical background that is strictly necessary to understand our research project on molecular similarity.

6.2 Molecular Representations

6.2.1 2D Representations

Before performing any kind of computer calculation on chemical entities, it is necessary to represent such chemical entities in a format suitable for digital storage and retrieval. Many organisations maintain databases of chemical compounds. Some of these databases are publicly accessible, others are proprietary. Among the most popular publicly accessible databases, we wish to mention DrugBank [89]–[93], ChEMBL [94]–[96], and the Protein Data Bank (PDB) [97]–[99].

An obvious way to digitally store chemical structures would be to save pictures of the 2D molecular graphs (Figure 6.1). However, image files have little values for cheminformatics. More appropriate representations are needed to store the molecules in

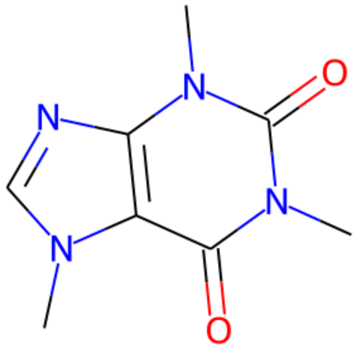


Figure 6.1: Picture of a 2D molecular graph. The writer's favorite molecule is represented: caffeine.

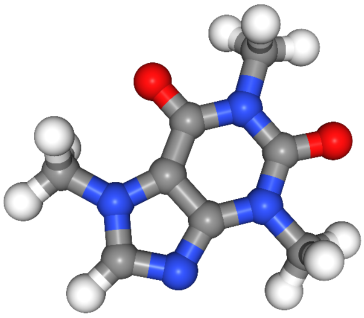


Figure 6.2: Picture of a 3D structure of caffeine.

databases for subsequent retrieval, and to perform calculations on the chemical structure. The most common format to store molecular graphs for cheminformatics applications is the Simplified Molecular-Input Line-Entry System (SMILES) [100]. SMILES encode 2D representations of molecules as linear strings of alpha-numeric characters. The SMILES for caffeine (the molecule represented in Figure 6.1) is: "`Cn1c(=O)c2c(ncn2C)n(C)c1=O`". Hydrogens are usually omitted (considered implicit) in SMILES strings and 2D pictures.

In their simplicity, SMILES encode for chemical information such as atoms, their chemical character, bonding patterns, branch points, and the presence of stereo centers. For a given molecule, there may be many valid SMILES strings. The process of generating a SMILES string for a molecule is not unique. In practice, the basic SMILES are not used for database storage. SMILES strings are often ordered in a unique way, for easier database retrieval and querying. The process of giving the SMILES a unique order is called canonicalization [101], [102].

In the present work we use the SMILES format as the main input for all cheminformatics calculations.

6.2.2 3D Representations

The 2D molecular representations only which atoms are bonded together. The 2D molecular representations are very useful for database storage and retrieval, for similarity analysis, and for the calculation of some chemical properties. In reality, molecules are 3D objects, and their steric and electronic properties depend on how atoms can be positioned in space to produce 3D structures (also called conformations). The 3D structures can be saved as image files (Figure 6.2). For bigger molecules, a single image would not be enough to view all the chemical components. Multiple images, captured from different angles, should be used. Even better, a 3D structure should be viewed with an interactive tool that allows users to move and rotate the conformer.

The most basic format for storage of 3D structures is the XYZ file format (extension

.xyz). The first line contains a number: the total number of atoms in the molecule. Then, there is an empty line. Each of the other lines represent an atom. Each atom line has four columns. The first column on the left is the atomic element. The remaining three columns are the X, Y, and Z coordinates. The contents of the XYZ file for a conformer of caffeine are reported in Figure 6.3.

```
24

C     -1.369306    2.818665    0.349239
N     -0.977497    1.435396    0.154315
C      0.326020    1.130840    0.067756
O      1.161996    2.059551    0.159214
C      0.811229   -0.141164   -0.115963
C     -0.117570   -1.163186   -0.216297
N      0.552330   -2.321669   -0.391800
C      1.884585   -2.065709   -0.406103
N      2.006830   -0.739997   -0.237137
C      3.274566   -0.047787   -0.192085
N     -1.432538   -0.841061   -0.127033
C     -2.397407   -1.906437   -0.231746
C     -1.856229    0.424238    0.053605
O     -3.088474    0.666073    0.129406
H     -0.616382    3.303244    1.000530
H     -1.320561    3.296620   -0.672150
H     -2.399802    2.924850    0.718100
H      2.661144   -2.820249   -0.532607
H      3.359321    0.509545    0.765991
H      4.129723   -0.741731   -0.292757
H      3.324808    0.749781   -0.990868
H     -1.928107   -2.852035    0.112995
H     -2.665915   -1.992301   -1.319857
H     -3.322765   -1.685476    0.340944
```

Figure 6.3: The 3D structure of caffeine in XYZ format.

There are more complex file formats for 3D structures. Some file formats can be used to store multiple conformations of a given molecule, or even multiple molecules with multiple conformations.

In the present work, we did not have a preference for any particular file format. We mostly used the PDB [103] and MOL2 [104] formats for storage of 3D structures, because they are the most common. We used the formats required by each specific software tool, and we converted between each format when needed (paying attention to preserve all chemical information).

6.3 Conformer generation

The problem of 3D molecular representation is challenging. Most molecules of interest can adopt more than one low-energy conformation. For bigger molecules, the number of accessible structures is very large. It is therefore necessary to take conformational flexibility into account. This usually means generating a set of conformations for a given molecule. The generated conformations should be a representative sample of all the low-energy conformations of the molecule. In cheminformatics, the most important conformations are the bioactive ones.

Many conformer generation tools and techniques are available [105], [106]. In the present work, we have used OMEGA [107]. OMEGA is a popular choice for conformer generation: its ability to produce realistic structures has been thoroughly tested [108]–[111]. OMEGA was designed to provide a representative sample of the conformational space of druglike molecules. The OMEGA algorithm is divided in four steps:

1. Preparation of a fragment database: a large collection of commercialy available compounds (in their 2D representations) is fragmented into continuous ring systems and small linear linkers. One or more 3D conformations are generated for each fragment.

2. Generation of torsion library: by analysis of a set of experimental crystal structures (mostly from the PDB), a set of torsion rules is generated (with associated common angles), in order to match every rotatable bond in small molecules with at least a torsion.

3. Structure generation: an input 2D graph is fragmented in the same way as the fragment database, and the fragments are reassembled into the parent molecule using geometric and chemical rules.

4. Torsion driving: the rotatable bonds of the conformer generated in step 3 are compared to the torsion library generated in step 2. The appropriate angle values for each torsion in the conformer are noted. All possible conformers are generated with all combinations of torsion angles.

5. Conformational sampling: step 4 generated a large amount of conformations, many of which have very high energy due to internal clashes. Many of the con-

formers from step 4 are redundant, since are similar to each other. In this last step, the conformer energies are evaluated using the MMFF94 force-field [112]. High energy conformers (due to internal clashes) are discarded. Similar conformers (based on an RMSD cutoff) are discarded (only the lowest energy conformer in a set of similar conformers is kept).

The OMEGA algorithm is available as a command line software tool by OpenEye Scientific [113].

6.4 Molecular similarity

Molecular similarity is a fundamental concept in cheminformatics [114]–[116]. It has wide applications in chemical database searching [117], [118], and medicinal chemistry [119], [120]. Molecular similarity applications are based on the so-called "Similarity Principle": similar molecules have similar properties and activities [121].

The Similarity Principle is of course an oversimplification. It is true only in a statistical sense: in a given set of compounds, molecule-pairs with high calculated similarity have, on average, more similar properties and activities than molecule-pairs selected at random [122]. There are known situations where molecule-pairs with high similarity measures exhibit very different activities [123]. It should be noted that the reverse of the Similarity Principle is not necessarily true: molecules that exhibit similar activities could be very dissimilar.

The Similarity Principle is nonetheless very useful, and molecular similarity has many practical applications.

6.4.1 2D similarity

Molecular similarity is often a function of the 2D molecular graphs, and it is calculated on molecular fingerprints. The most common types of fingerprints are fixed-size arrays of ones and zeros (i.e., binary arrays). For instance, two molecules A and B could have fingerprints $F_A = [1, 0, 1, 1, 0]$ and $F_B = [0, 0, 0, 1, 1]$. Each element of the fingerprints encodes the presence (1) or absence (0) of a chemical feature. Many functions can be used to calculate similarity on pairs of fingerprints [124]. The most common similarity measure is the Tanimoto coefficient [125], presented in Equation 6.1.

$$T_{A,B} = \frac{N_{A,B}}{N_A + N_B - N_{A,B}} \tag{6.1}$$

N_A and N_B are the number of 1 elements in the fingerprints of molecules A and B, respectively. $N_{A,B}$ is the number of 1 elements that are present in both molecules A and B. In our example, $N_A = 3$, and $N_B = 2$. $N_{A,B} = 1$, since there is only a single 1 element in the same position in both fingerprints (the fourth element). So, $T_{A,B} = \frac{1}{3+2-1} = 0.25$ is the Tanimoto coefficient for molecules A and B. The Tanimoto coefficient can have values between 0 and 1, with higher values for more similar molecules.

The interpretation of the Tanimoto coefficient depend on the type of fingerprint that was used on the molecules. There are many algorithms for fingerprint generation [124]. Each algorithm parses the 2D molecular graphs in a specific way, and encodes specific types of chemical features in the elements of the fingerprint array. The Tanimoto coefficients represent molecular similarity with respect to the types of chemical features that are encoded in the fingerprints.

The 2D similarity measure used in the present work is the Tanimoto coefficient calculated on CDK Extended fingerprints [126]–[128]. CDK Extended fingerprints are hashed fingerprints [117], a type of fingerprints that were first developed by Daylight Chemical Information Systems [129]. Algorithms that generate hashed fingerprints calculate unique linear paths through a molecular graph, and encode them in the elements of the fingerprint array by applying a hash function. In most hashed fingerprints, only linear paths are considered. The CDK Extended fingerprint also considers ring systems.

6.4.2 3D similarity

The most common example of 3D similarity is pharmacophore similarity [117]. A pharmacophore is a set of chemical features with their relative 3D spatial orientation. Usually, pharmacophores use a very broad and generic definition of chemical features (e.g., aromatic rings, ions, hydrogen bond donors and acceptors...).

In the present work, we have used the ROCS algorithm [130] for 3D similarity calculations. ROCS considers pharmacophore similarity and molecular shape similarity. Actually, the ROCS algorithm was first developed to only consider shape similarity [131], pharmacophore similarity was included later.

Before considering shape similarity, it is necessary to understand what a volume is. A volume is the integral of scalar field Equation 6.2. A scalar field is a function that has a single number value (a scalar) at any point in space.

$$V = \int f(x, y, z) dV \tag{6.2}$$

The scalar field f is also called the characteristic function of the volume V. For the common understanding of volume (the “size” of an object), f has value 1 at any point inside the object, and value 0 outside. But the scalar field f could also have different values, and we could still calculate the volume V. A volume is a contraction of the information represented by the scalar field. For instance, two objects must have the same volume in order to have the same shape, but two objects with the same volume can have a different shape.

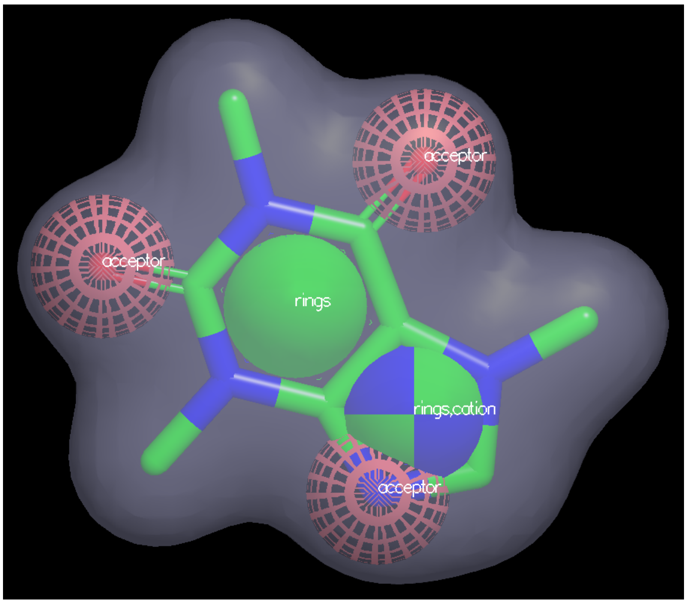


Figure 6.4: ROCS shape and pharmacophore surfaces of caffeine, generated with vROCS.

Given the definition of volume, an obvious measure of shape similarity (or more precisely, shape distance), would be:

$$D = \int [f(x, y, z) - g(x, y, z)]^2 \mathrm{d}V \tag{6.3}$$

f and g are the scalar fields of two objects. The distance D has higher values for objects with very different shape, and has value 0 for objects with identical shape. Any distance measure D can be converted to a similarity measure S [125]:

$$S = \frac{1}{1 + D} \tag{6.4}$$

Similarity measures have higher values for more similar objects. Equation 6.3 can also be expressed as:

$$\begin{aligned} D^2 &= \int f(x, y, z)^2 \mathrm{d}V + \int g(x, y, z)^2 \mathrm{d}V - 2 \int f(x, y, z) g(x, y, z) \mathrm{d}V \\ &= I_f + I_g - 2O_{f,g} \end{aligned} \tag{6.5}$$

This is the fundamental equation for shape comparison. It can be used for any kind of scalar fields, not only the scalar fields corresponding to the common understanding of “volume”. The terms I_f and I_g are the self-volume overlaps of fields f and g, respectively. $O_{f,g}$ is the overlap between the scalar fields of the two objects. The I terms are independent of orientation. On the other hand, $O_{f,g}$ depends on the relative orientation of the two objects. Maximizing $O_{f,g}$ (thus minimizing the square distance D^2) is equivalent to finding the best overlay between the two objects. If the two objects are molecules, this means aligning them by rotating and translating them, without changing the internal coordinates (that specify the molecular shape).

I_f, I_g, and $O_{f,g}$ can also be used to calculate other types of similarity measures. The Tanimoto coefficient in 3D space is:

$$T_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}} \tag{6.6}$$

Note the resemblance to the 2D formulation of the Tanimoto coefficient (Equation 6.1).

So far, the mathematical formulation has been pretty straightforward. Molecules would be represented as a set of spheres (one sphere for each atom), with a scalar field having value 1 inside the sphere, and 0 outside. But the computational implementation is difficult and not very efficient. The difficulty arises from the fact that spheres of different atoms can overlap with each other: the intersection volume must not be counted multiple times! The computation becomes increasingly more complex and slow with bigger molecules.
A breakthrough came with the work of Grant and Pickup [132]. They showed that if one let go of the concept of the scalar field being binary, and instead use a sum of continuous functions, the sphere volume could be recovered with high accuracy. Radial Gaussian functions were the function of choice:

$$f(r) = e^{-wr^2} \tag{6.7}$$

The main advantage of Gaussian functions is that the overlap of two atomic Gaussians produce another Gaussian function. The speed of the algorithm greatly increased by using Gaussians instead of spheres to calculate molecular volumes and to perform alignments.

As mentioned earlier, ROCS do not only consider shape similarity, but also pharmacophore similarity (called "color similarity" in the ROCS documentation). The default chemical features (color features) considered by ROCS are: hydrogen bond donor and acceptor, anion and cation, hydrophobic group, and ring. Other chemical features could be specified by the user. The color features are represented by radial Gaussians, just has the atoms during the shape similarity calculation.

The ROCS algorithm is available as a command line software tool by OpenEye Scientific [133]. There is also a graphical software that implements the ROCS algorithm, vROCS. With vROCS, users can visualize the molecular shapes and pharmacophores, and the alignments (Figure 6.4).

6.5 Computational Models

An important goal of cheminformatics is the creation of models that relate structural features of molecules (descriptors) to their biological activity or to their physico-chemical properties [134], [135]. This subfield of cheminformatics is named Quantitative Structure-Activity Relationship (if a biological activity is being modeled) or Quantitative Structure-Property Relationship (if a physico-chemical property is being modeled), and it is abbreviated QSAR or QSPR. The term QSAR is more general, and we will be using it in the rest of this work.

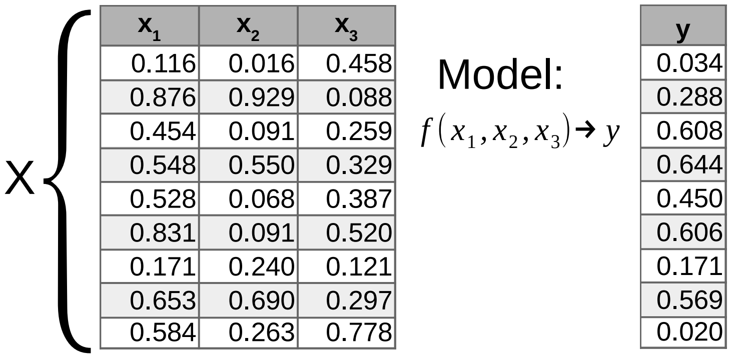
The scope of QSAR models is both theoretical and practical. From a theoretical perspective, QSAR models contribute to a better understanding of the structural features that result in interesting chemical activities. From a practical perspective, QSAR models can be used to predict the activity of compounds for which an activity has never been measured. Biological activities are usually expensive and complex to measure experimentally. On the other hand, the chemical descriptors used as input for the QSAR models are usually very cheap and easy to obtain. They are usually calculated by computers on 2D (but also 3D) molecular representations. QSAR models could even be used to predict the activity of molecules that have never been synthesized.

The quality of a QSAR model is highly dependent on the data-set that was used to build the model (i.e., the training-set). The prediction of the activity of molecules outside the training-set is reliable only if the new molecules are similar to compounds included in the training-set. Even if this criterion is met, QSAR models may fail [123]. This problem is partly due to the excessive simplicity and inflexibility of the models being used, or to the lack of proper techniques to evaluate the model performance [136], [137].

It should be noted that the modeling tasks that are described in the present work are different from the typical QSAR problem. In the typical QSAR problem, the data-set consists of a list of molecules. Each row of the data-set represents a single molecule, each column represents a molecular descriptor. There is also a column for the "target variable": the activity being model (biological or physico-chemical). The model relates the molecular descriptor columns to the target variable (Figure 6.5). Models are usually functions f that take as input a vector $\mathbf{x}$ (whose elements are the descriptors x_1, x_2...) and calculate an activity y. The model function f depends on a set of parameters, and the parameters affect the calculated y. Simple models (e.g., linear models) usually depend on few parameters, whereas very complex models (e.g., neural networks) depend on many parameters. Models are built with an optimization algorithm, that minimizes the error between the y calculated by the model and the true y in the training-set. This is accomplished by searching for the best set of model parameters.

The computational models that we built in the present work are different from the typical QSAR models.

Typical QSAR models are regression models: the target variable y is a real-valued num-



x_1	x_2	x_3
0.116	0.016	0.458
0.876	0.929	0.088
0.454	0.091	0.259
0.548	0.550	0.329
0.528	0.068	0.387
0.831	0.091	0.520
0.171	0.240	0.121
0.653	0.690	0.297
0.584	0.263	0.778

y
0.034
0.288
0.608
0.644
0.450
0.606
0.171
0.569
0.020

Figure 6.5: General data modeling process. The three columns x_1, x_2, and x_3 are the input features of a model f, that is built to predict the target variable y. In QSAR, the input features are molecular descriptors, and the target variable is an experimentally measured biological activity (or physico-chemical property). Each row contains the descriptors and the activity of a single molecule.

ber (e.g., 3.2 or 0.58). The task that we wanted to accomplish with computational models (similarity-prediction, see chapter 7) is a classification task. So, we built classification models: the target variable y is categorical (e.g., Male-Female, Yes-No, Car-Truck-Bike). A real-valued variable can be easily converted into a categorical variable by applying thresholds. For instance, the real-valued variable "age of a person in years" can be converted into the categories Child-Adult-Elderly by applying arbitrary thresholds. Converting a real-valued target variable into a categorical target variable is sometimes very helpful: well-defined categories are more easily interpretable by humans than numbers.

The other important difference between the models described in the present work and the typical QSAR models is the input features. As we explained earlier, QSAR models typically use molecular descriptors as input features, and each row contains the descriptors for a single molecule. On the other hand, each row of the data-sets we have used represents two molecules (a molecule-pair). The input features are molecular similarity measures. Each row contains the similarity measures for a molecule-pair. This choice of input features (and of data-set construction) was made because of the particular task that we wanted to accomplish (similarity-prediction).

Among the many types of classification models, we choose to use Logistic Regression (LogReg), Decision Tree (DT), and Random Forest (RF) models [138]–[142]. We wanted to explore the ability of simple and relatively inflexible models such as LogReg, and of more complex and flexible models such as DT. Simple models tend to underfit the training-set: they are not able to represent complex patterns in the data. Complex models tend to overfit the training-set: they memorize the noise in the data alongside the meaningful patterns. The distinction between simple and complex models is a well-known problem, called the bias-variance tradeoff (Figure 6.6) in the literature [143]. So, we choose LogReg and DT models because they are at the extremes of the bias-variance spectra. We also used RF models: they are designed to reduce the overfitting of DTs, and are usually a good compromise between bias and variance.

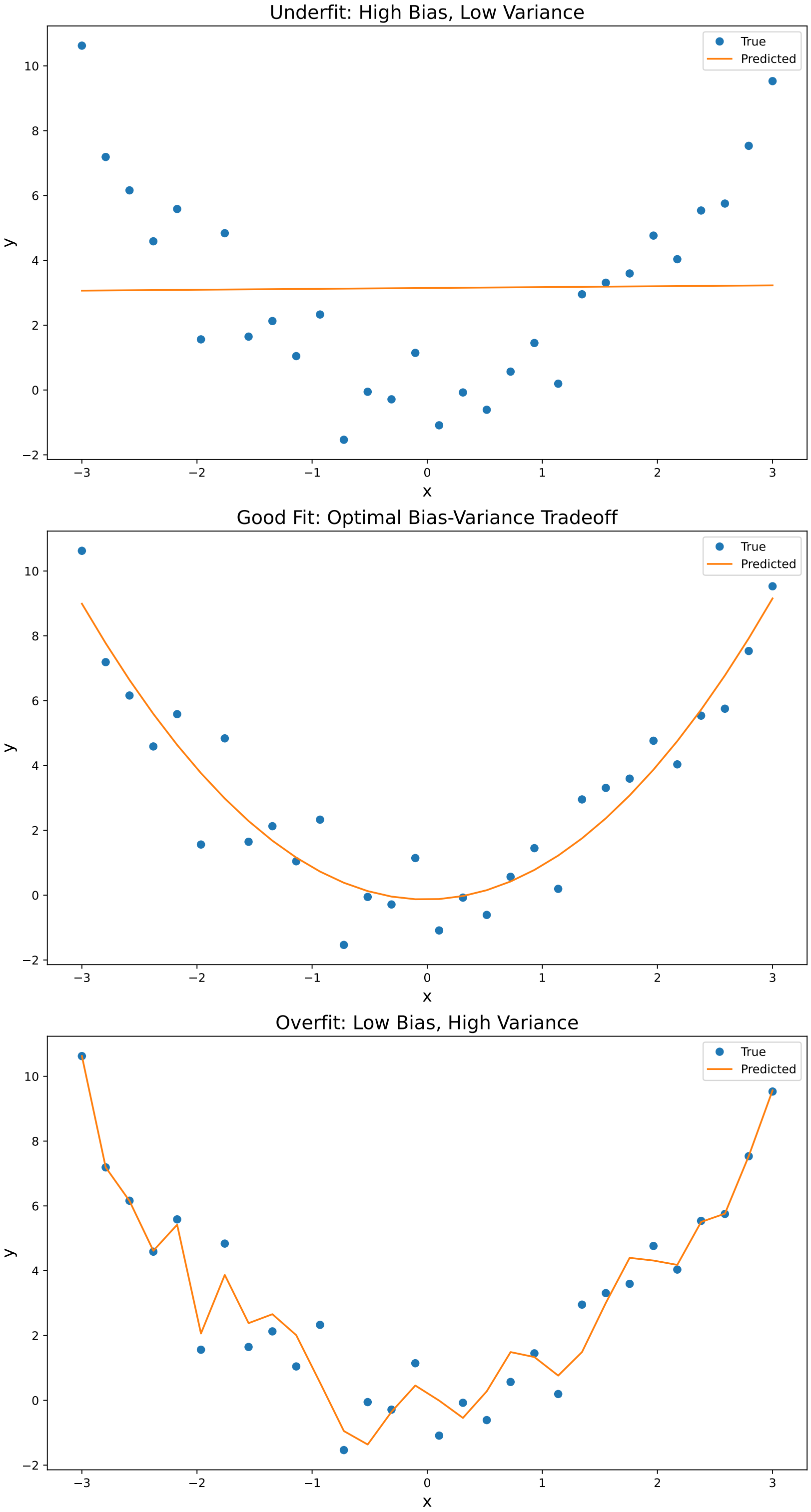


Figure 6.6: Representation of the bias-variance tradeoff problem.

Chapter 7

Introduction

An orphan drug is a medicinal product used to treat a rare disease that affects only a small number of patients (the actual number of patients depends on the local legislations) [144]. Given the small number of patients affected by the rare disease, and the high costs involved in modern drug discovery programs [145], [146], orphan drugs are not an immediately attractive market for pharmaceutical companies.

To encourage pharmaceutical companies to develop orphan drugs, regulatory agencies have brought forward legislations that provide a range of incentives. Such incentives include grants, financial incentives, the possibility of an accelerated review, and market exclusivity. Market exclusivity is arguably the most important incentive: under the EU legislation, a pharmaceutical company that develops an orphan drug for a specific rare disease is given a 10-year period of market exclusivity. During this period no products that are considered similar to that orphan drug can be accepted or authorized by any European regulatory competent authority. Orphan drugs have less competition than conventional drugs, which encourages pharmaceutical companies to invest in researching novel medicines for rare diseases.

Currently, there are many ways to define whether two molecules are similar or dissimilar. The assessment of similarity between two drugs takes into account three criteria: molecular structure, mechanism of action, and therapeutic indication. Two drugs will be considered diverse if there are significant differences in one or more of the three aforementioned criteria. Thus far, the European Medicines Agency (EMA) has used majority voting on discretional judgments of similarity when assessing new drugs for rare diseases. Similarity is an inherently subjective concept, which depends on individual factors such as gender, age, state of mind, and previous experiences [147], [148]. In general, chemical structure information is perceived differently by different individuals [149], but a fair level of consistency can be achieved using a wisdom of crowds approach [150].

Automated procedures that quantitatively and objectively evaluate molecular similarity are needed. Such an algorithm would not replace the current human-based processes used to evaluate applications for orphan drugs authorizations. Instead, it would produce an useful quantitative input to be considered by the human experts evaluating the application. Franco *et al.* have developed a computational procedure that calculates the probability that a pair of molecules would be considered similar by a crowd of experts [151], [152]. The procedure is based on Logistic Regression (LogReg) models [140], [141]. LogReg models are trained on Tanimoto coefficients calculated on different 2D molecular fingerprints. The procedure successfully reproduced human assessments of molecular similarity, both on the data set used to train the LogReg models (the training-set), and on an external test-set.

From now on we will call "similarity-prediction" the computational prediction of human assessments of molecular similarity. We will call "similarity-prediction models" all kinds of models that perform similarity-prediction.

The procedure was simple and effective, but it has room for improvements, as the authors of the original work themselves suggested. For instance, the Tanimoto similarity could be calculated on other kinds of molecular representations, which encoded 3D structural information of the molecules. This possibility was explored by Franco *et al.* in their second work on the subject [152]: the LogReg models were trained on proprietary 3D molecular fingerprints by MOE [153]. Such 3D molecular fingerprints performed worse than the simpler 2D counterparts. That type of 3D molecular fingerprints compresses the 3D structural information to a fingerprint, which is a 1D vector. This may be the cause of the ineffectiveness of the 3D molecular fingerprints for this particular application.

Other more advanced techniques can be used to encode 3D structural information of a molecule, and to calculate similarity between a pair of molecules. Such techniques should encode important structural information of the molecules, such as their shape and the spatial orientation of pharmacophoric groups. A perfect example of similarity measure which takes into account molecular shape and orientation of pharmacophoric groups is TanimotoCombo by ROCS [130], [133], [154]. Contrary to simple 3D fingerprints, ROCS do not compress 3D molecular information: the information is held in a 3D numerical tensor, and the similarity measure is calculated on a pair of such 3D tensors.

A different approach would be that of VolSurf descriptors [155], [156]. The VolSurf approach is based on Molecular Interaction Fields (MIFs). A MIF is a 3D tensor that describes the spatial variation of the interaction energy between a molecule and a chosen probe [157], [158]. VolSurf uses MIFs to calculate a set of relevant physico-chemical properties. Similarity measures can be calculated on vectors of VolSurf properties.

For these reasons, ROCS and VolSurf can be more effective in capturing 3D molecular similarity than simple 3D fingerprints. They can be used to complement information provided by 2D fingerprints, and to possibly improve models that predict the outcomes of similarity voting.

The aim of this project is not only to develop new computational models based on 3D molecular similarity, but to investigate the decision-making process of human experts asked to assess the similarity of a pair of molecules. Do human experts only take into consideration the 2D molecular graph, or do they also consider the 3D conformers? What about difficult and borderline cases, where 2D and 3D molecular similarity measures do not agree? Will experts agree with each other? It is our belief that a better understanding of the human decision making process will provide important information to agencies that rely on human similarity judgments for the assessment of orphan drug status. It will also help in developing new algorithmic tools that support human experts, to provide clearer and less biased judgements. Such tools could also be used by pharmaceutical companies to perform preliminary virtual screenings of molecules that are suitable to receive the orphan drug status.

Chapter 8

Methods

8.1 The original training-set

The first set of models that we developed is based on the training-set originally created by Franco *et al.* [151]. They selected 1068 drug-like molecules from DrugBank 3.0 [91], and computed ECFP4 fingerprints [159] on each molecule. They calculated Tanimoto similarity on each pair of molecular fingerprints. They then selected 100 pairs of molecules, which covered the widest and most uniform spread of Tanimoto values. This set of 100 pairs of molecules was evaluated by several individuals involved in European, American, Taiwanese, and Japanese regulatory authorities. In total, 143 experts evaluated the 100 pairs of molecules. The experts were asked to evaluate whether each pair of molecules was composed by similar (Yes) or dissimilar (No) molecules. Franco *et al.* collected the expert evaluations, and for each pair of molecules, calculated the fraction of experts which considered the pair to be similar or dissimilar. If the fraction of expert that considered a pair of molecules to be similar was ≥ 0.5, Franco *et al.* considered the pair to be similar, otherwise the pair was considered dissimilar. This training-set of 100 pairs of molecules, accompanied by expert evaluations, was kindly made publicly available (Table S1 of article [151]).

Franco *et al.* calculated a variety of fingerprints on each molecule in the training-set. They then calculated Tanimoto coefficients based on all fingerprints for all the 100 pairs of molecules. They used the Tanimoto coefficients to build similarity-prediction LogReg models.

8.2 The original test-set

Franco *et al.* tested the LogReg models created with the training-set on an external test-set. This data-set was confidentially provided to Franco *et al.* by EMA's Committee for Medicinal Products for Human Use (CHMP). It consisted of 100 pairs of molecules.

Each pair of molecules was composed by an existing orphan drug for a specific rare disease, and by another molecule that was submitted for approval as a treatment for the same rare disease. Each pair of molecules was accompanied by the CHMP evaluation of new molecule proposed for the treatment.

The molecules included in the test-set are quite different from the ones in the training-set [151]. Compounds in the test-set are significantly larger than compounds in the training-set. But most importantly, of the 100 molecule-pairs in the test-set, 89 of them had been judged to be non-similar pairs with only 11 judged to be similar pairs, whereas the training-set contained near-equal numbers of the two types of molecule-pair.

The use of an external test-set to evaluate prediction models is universally considered a best practice [137], [139]. We asked CHMP to provide us confidential access for the original test-set used by Franco *et al.* EMA's CHMP kindly approved our request, but at the time of writing we have not yet received the whole data-set. We will not be able to evaluate our models in the same way that Franco *et al.* did.

To follow the best practices and evaluate the models on an unseen data-set, we will employ a procedure that we will call cross-test. We will use the new data-set (see below sections about the creation of the new data-set) to evaluate models built on the original training-set. Vice versa, we will use the original training-set to evaluate models built on the new data-set.

Cross-testing is a valid procedure to evaluate similarity-prediction models. The original training-set as well as the new data-set were specifically built to include a diverse set of molecules, and to represent a wide spectrum of molecular-similarity instances. So using one data-set to test models built on the other data-set is a good way to evaluate the performance on difficult and interesting molecule-pairs.

8.3 Modeling the original training-set

8.3.1 The 2D protocol

We reproduced similarity-prediction models based on open source 2D fingerprints that were described by Franco *et al.* [152]. We will call these models "2D similarity-prediction models".

The protocol to build 2D similarity-prediction models begins with the preprocessing of original SMILES using RDKit [160] and MolVS [161]. SMILES strings are standardized. Then, the counterions are removed, and the remaining species neutralized. We then validated all preprocessed molecules with MolVS, and visually checked them. We computed all 2D fingerprints available in RDKit and CDK [126], [127], and calculated Tanimoto coefficients on each pair of molecules, with each 2D fingerprint.

8.3.2 The MOE 3D protocol

We developed a new 3D protocol based on MOE software, that uses force field MMFF94x, and cosine similarity calculated on vectors of standardized VolSurf descriptors [155], [156]. VolSurf descriptors were chosen because they convert MIFs in relevant physical-chemical properties. VolSurf descriptors are independent of conformational sampling and alignment [155], resulting in a more straightforward protocol. The protocol starts with SMILES preprocessing using the Wash command, with a neutral protonation state, and by preserving chirality information. We then generated a minimized conformer for each molecule using the Energy Minimize command, with chirality preservation and optimal orientation of OH groups. We calculated VolSurf descriptors for each unique molecule in the dataset. We then applied standard scaling on each VolSurf descriptor, using all the unique molecules in the dataset. Cosine similarity between each pair of molecules is then calculated on vectors of standardized VolSurf descriptors: we will call this value "VolSurf Similarity".

8.3.3 The OpenEye 3D protocol

We developed another 3D protocol based on OpenEye software. The protocol starts with SMILES preprocessing using Filter command (included in OMEGA software [113]) with an empty filter file (so molecules are just preprocessed, but no molecule is discarded). We use OMEGA classic algorithm to generate up to 200 conformers for each molecule. Conformers generated by OMEGA are ready to use, since OMEGA was designed to sample the conformational space around solid-state structures of drug-like molecules [107]. For each pair of molecules, we use ROCS to perform all possible conformer alignments, and to calculate similarity scores for each alignment. Then, we keep the TanimotoCombo score corresponding to the best alignment, for each pair of molecules.

8.3.4 Training of similarity-prediction models

We used scikit-learn [71] to build similarity-prediction models using the original training-set. We used Tanimoto coefficients calculated on all available open source fingerprints as input features for the 2D similarity-prediction models. We will focus on models built using CDK Extended fingerprint [126]–[128], which was considered the best fingerprint by Franco *et al.* [152].

We used cosine similarity calculated on vectors of standardized VolSurf descriptors as input feature for similarity-prediction models for the MOE 3D protocol. For OpenEye 3D protocol, the input feature of similarity-prediction models is the TanimotoCombo score of best conformer alignment for each pair of molecules.

Using the aforementioned input features for the 2D and 3D protocols, we built LogReg similarity-prediction models. From now on, we will call "single-feature models" the similarity-prediction models built using just one input feature from one protocol. The 2D model based on CDK Extended fingerprint and the OpenEye 3D model based on TanimotoCombo were successful in the similarity-prediction task on the original training-set: they successfully reproduced human assessments of molecular similarity (see Results Chapter). On the other hand, the model built using the input feature from the MOE 3D protocol showed poor similarity-prediction power. The MOE 3D protocol was not considered for further modeling.

We then built models based on the two most promising input features: the Tanimoto coefficient calculated on CDK Extended fingerprint (Tanimoto CDK Extended), and TanimotoCombo. We will call these models double-feature models. They combine the predictive power of a 2D and a 3D feature. We built LogReg double-feature models, introducing L1 and L2 regularization. We also built more complex double-feature models: Decision Tree (DT), and Random Forest (RF) models [138]–[142].
We built two sets of double-feature models. A first set of double-feature models were built with default hyperparameters from scikit-learn (default double-feature models). For the second set of double-feature models (tuned double-feature models), we fine-tuned the hyperparameters with grid search using 10-fold cross-validation (CV) [140], [162].

The hyperparameter tuning was performed in order to reduce overfitting of the models. For the basic LogReg model, there was no hyperparameter to tune. For LogReg models with L1 and L2 regularization, we tuned the regularization strength hyperparameter. For DT, we tuned the maximum depth of the tree, the minimum number of samples required to split an internal node, and the minimum number of samples necessary to be at a leaf node. For RF, we tuned the same hyperparameters that were tuned for DT, and additionally the number of trees in the forest.

8.3.5 Model Performance Evaluation

We evaluated the similarity-prediction models using a variety of performance metrics for classification problems [163]. The most easy-to-understand metric is the number of samples that a model correctly classifies ($N_{correct}$). Since we will be focusing on data-sets that are composed by 100 samples (the original training-set and the new data-set, but also the original confidential test-set that we have not yet received), $N_{correct}$ can be considered the percentage of correctly classified samples (a metric that is usually called "Accuracy"). But such a simple metric, although easily understandable and interpretable, has many limitations [163]. For instance, it does not take into consideration the class probability calculated by the model, and the fact that different probability thresholds can be used.

To overcome these limitations of the $N_{correct}$ metric, we also used other popular metrics for performance evaluation: the Area Under the Receiver Operating Characteristic curve (AUROC), the Average Precision (AveP), the Log Loss (L_{log}), and the Brier score (L_{Brier}). AUROC and AveP values are between 0 and 1, the higher the better. L_{log} and L_{Brier}, on the other hand, are "loss metrics", meaning that a better model will have a lower value. These metrics are better than N_{correct} in representing the overall performance of a classifier.

8.4 Creation of new data-set

We created a new data-set of human assessments of molecular similarity, to train new similarity-prediction models. We wanted to include a diverse set of molecules with known bioactivity, but that not necessarily possessed drug-like properties. So we decided to focus on molecules that targeted three well-known biological targets: HERG [164], 5HT2B [165], and CYP2D6 [166]. HERG and 5HT2B are anti-targets, whereas CYP2D6 is a liver metabolic protein. These receptors are well-known in the medicinal chemistry community, and bioactivity data is abundant. To make sure that all molecules had known bioactivity, we included only molecules for which an inhibition constant K_i [167] was measured. The creation of the new data-set consisted of two parts: the initial selection of molecule-pairs to be included, and the assessment of their similarity by experts through an online survey.

8.4.1 Selection of molecule-pairs

We queried ChEMBL 27 database of bioactive compounds [94]–[96] for molecules targeting HERG, 5HT2B, or CYP2D6. We only selected molecules for which a K_i value was measured. We selected 1307 compounds that targeted HERG, 1299 compounds that targeted 5HT2B, and 155 compounds that targeted CYP2D6. These initial lists of compounds still included some duplicate entries for each target. We used RDKit to calculate InChiKey [168], [169] values for each compound, and considered identical all compounds that shared the same InChiKey. We then applied the following rules on all subsets of identical compounds (for each target separately):

- If there are more than two identical compounds, apply Dixon's test [170] on their pChEMBL values, remove outlier entries, and keep first compound in subset with median pChEMBL value.
- If there are two identical compounds calculate the absolute difference of their pChEMBL values.
    - If the absolute difference is ≥ 1, drop both compounds.

  - Otherwise, keep the first of the two compounds, with mean pChEMBL value.

pChEMBLis a standardized measure of bioactivity of the ChEMBL database, suitable for comparing database entries and performing outlier detection [95].
After removal of duplicate molecules and outliers with the aforementioned procedure, we were left with 1201, 1172, and 115 unique molecules that target HERG, 5HT2B, and CYP2D6 respectively. We calculated the absolute difference of pChEMBL values for each molecule pair. We will call this quantity "pChEMBL distance".

We applied the 2D protocol to all compounds of the three targets, calculating Tanimoto CDK Extended between each unique pair of molecules. Of all unique molecules, 2 5HT2B compounds and 2 CYP2D6 compounds did not pass the preprocessing step of the 2D protocol, and were excluded from the next steps.
We then applied the OpenEye 3D protocol. Few molecules did not pass the OMEGA conformer generation step of the protocol: we were left with 1198, 1168, and 111 unique molecules for target HERG, 5HT2B, and CYP2D6 respectively. We would have had to perform ROCS calculations on 717003, 681528, and 6105 molecule-pairs for the three targets respectively.
The ROCS step of the protocol would have taken a huge amount of time: we roughly estimated 30 thousand hours using a single CPU (longer than the duration of the writer's PhD scholarship!). So we randomly selected 3000 molecule-pairs for each target, and performed ROCS alignment and scoring between all conformers of each of the 9000 total molecule-pairs. This step took around 6 days using a single CPU.

We created an initial data-set with 9000 rows (one for each randomly selected molecule-pair subjected to the ROCS calculation). This data-set included Tanimoto CDK Extended and TanimotoCombo values for each row. To classify molecules as either similar or dissimilar in 2D and in 3D, we used an approach based on a similarity threshold with a small buffer region, similar to the one described by Ehrman *et al.* [171]. We classified a molecule-pair as similar in 2D if Tanimoto CDK Extended $\geq$ 0.7, and as similar in 3D if TanimotoCombo $\geq$ 1.4. Such thresholds are popularly used for the two similarity measures [172]–[174]. In order to avoid an extreme sensitivity to small molecular differences around the thresholds [171], we used a 0.05 and a 0.1 buffer region for Tanimoto CDK Extended and TanimotoCombo, respectively. So we classified a molecule-pair as dissimilar in 2D if Tanimoto CDK Extended $\leq$ 0.65, and as dissimilar in 3D if TanimotoCombo $\leq$ 1.3.

We then divided the 9000 molecule-pairs data-set in 4 subsets: pairs that are similar in 2D and in 3D (*sim2D,sim3D*), pairs that are similar in 2D and dissimilar in 3D (*sim2D,dis3D*), pairs that are dissimilar in 2D and similar in 3D (*dis2D,sim3D*), and pairs that are dissimilar in 2D and dissimilar in 3D (*dis2D,dis3D*). Of the original 9000 molecule-pairs, 177 were classified as *sim2D,sim3D*, 54 as *sim2D,dis3D*, 97 as *dis2D,sim3D*. Since the 9000 molecule-pairs were randomly selected, the vast majority of them are dissimilar in 2D and in 3D: 8540 were classified as *dis2D,dis3D*. The subsets contain 8868 molecule-pairs in total; 132 pairs fell in the buffer zones, and could not be categorized as either similar or dissimilar in 2D or 3D: they were excluded from further steps.

Subsets *sim2D,sim3D*, *sim2D,dis3D*, and *dis2D,sim3D* are small enough to be visually inspected. We visually inspected them with DataWarrior [175], and manually selected 25 molecule-pairs from each subset. Since subset *dis2D,dis3D* was too big for visual inspection, we randomly selected 25 pairs from it.

We thus obtained a data-set with 100 molecule-pairs, containing 25 pairs from each similarity subset. This data-set contains 50 simple molecular similarity instances, where both the 2D and the 3D similarity measures agree on the similarity or dissimilarity of a molecule-pair. The data-set also contains 50 complex molecular similarity instances, where the 2D and 3D similarity measures disagree.

8.4.2 The survey

The data-set with 100 molecule-pairs (belonging to four similarity subsets) was created to be subjected to similarity assessment by human experts. The original training-set by Franco *et al.* contained molecule-pairs selected on the basis of only a 2D similarity measure. Our data-set was created selecting molecules based on a 2D (Tanimoto CDK Extended) and a 3D (TanimotoCombo) similarity measures. So we included 3D representation of the molecules, alongside classical 2D graph representations, in order to let the experts consider both aspects in their assessments.

A static 2D molecular graph is sufficient to represent the 2D structure of a molecule. On the other hand, a static 3D picture would be insufficient, since some parts of a molecule would be hidden. We wanted the 3D representations to be interactive: human experts should be free to rotate, translate, and zoom the 3D representations, to observe all parts of a molecule, and to focus on the parts that they deem more important.

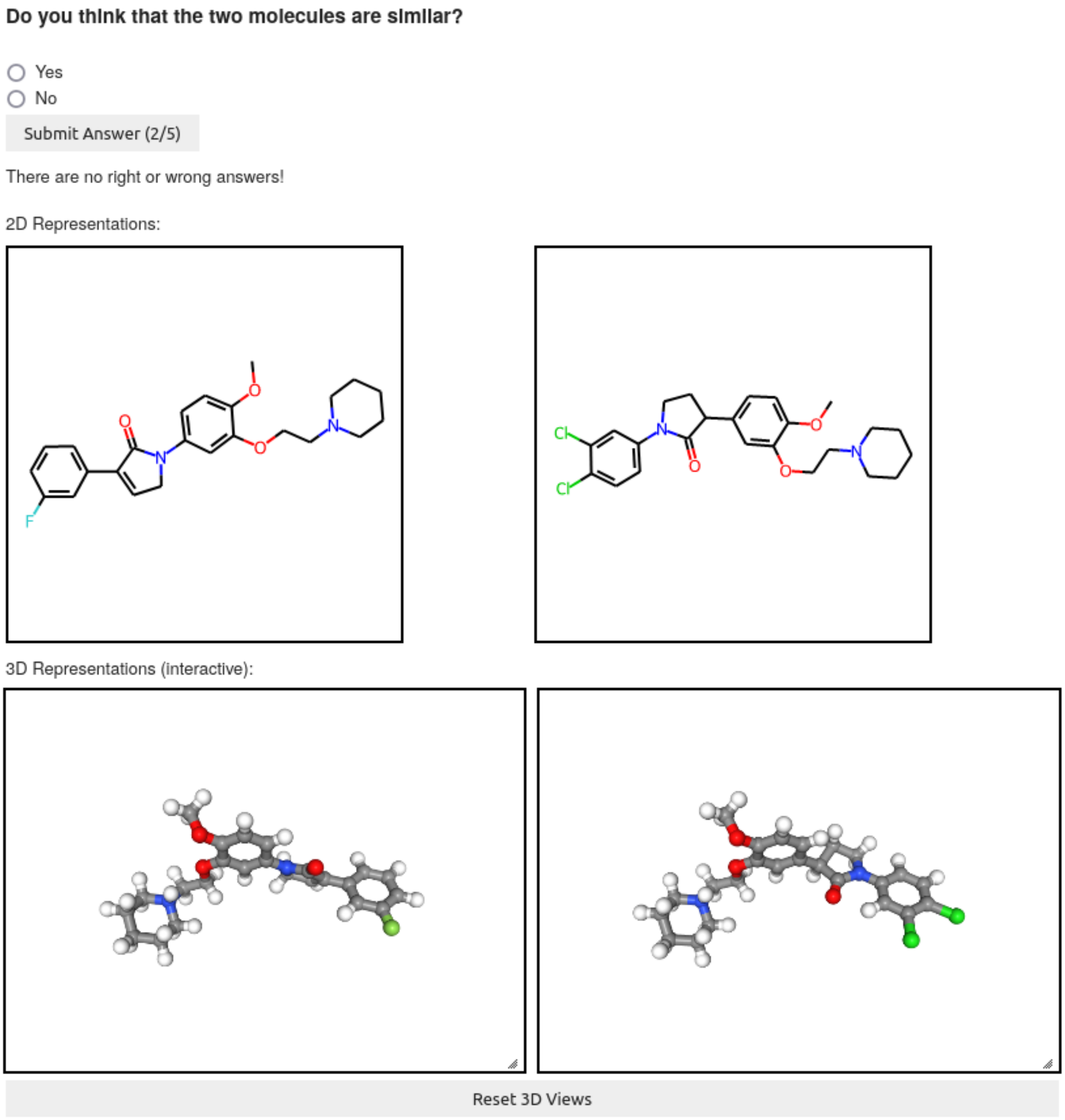
Franco *et al.* used MarvinSketch 5.5 to generate the 2D graph representations of the molecules that were subjected to similarity assessments by experts [176]. For each molecule-pair, the 2D graphs were not necessarily aligned to each other. The alignment can affect the similarity assessment of different experts: some experts would perform the alignment in their minds, and some other experts would not. When judging the similarity between two objects it is advisable to place them in a way that maximizes their overlap [147]. So we wanted the 2D graph representations, and the initial positions of the 3D interactive representations, to be aligned in a standard way that maximized the overlap between each molecule-pair, thus reducing the noise of the similarity assessments.

We used RDKit to calculate the Maximum Common Subgraph (MCS) of each molecule-pair. We then aligned both molecules in each pair to their MCS, and generated 2D graph representations after the alignment. The 2D graph representations were saved on disk

as SVG image files.
The ROCS step of the OpenEye 3D protocol produced, for each molecule-pair, a MOL2 file [104] containing the best alignment of OMEGA conformers. MOL2 files with conformations of two different molecules were not suitable to visualize two molecules side-by-side. And even though conformers in the ROCS output were aligned, they did not appear well-centered in the molecular visualization tool of choice, NGLview [177], [178]. For each molecule-pair, we computed the transformation matrix required to align the principal axes of the biggest molecule to the X, Y, Z axes. We then applied the transformation matrix to both conformers of the molecule-pair. This transformation preserved the ROCS alignments, made the conformers appear well-centered in the NGLview windows. We then saved the conformers in two separate PDB files [103] for each molecule-pair.

An online survey was the best way to obtain expert similarity assessments on molecular representations that followed the aforementioned principles. We programmed a web application that delivered the survey. The application was created with Voilà [179], a tool to convert Jupyter notebooks [180] in standalone web applications. The application started by asking users to recognize a randomly generated word (a simple captcha problem), to make sure that the survey was not attacked by malicious actors. Users were then presented the conditions of the survey: that they would be presented 5 pairs of molecules, and that they would be asked to judge whether they considered each pair to be similar or dissimilar. They were also informed that the survey should take about 5 minutes to complete, that there were no right or wrong answers, and that their anonymity was guaranteed. Users that accepted the conditions were immediately presented the first molecule-pair. For each user, we randomly selected 5 molecule-pairs. Users could be affected in their assessments by the relative positions of the 2D and 3D representation. We did not want the results to be biased by this fact, so each users was either shown the 2D representations above the 3D representations of a molecule-pair, or the other way around. Users were randomly assigned to one of the two treatments. An example of the appearence of the survey is shown in Figure 8.1. Users were free to interact with the 3D representations, generated by the application as NGLView Jupyter widgets. Users were free to return to the initial position of the 3D representation by clicking a "Reset 3D Views" button, below the 3D representations. Users had to express a similarity assessment for each of the 5 molecule-pairs that they were presented. The application did not allow users to proceed in the survey without answering. After users answered the 5 questions about molecular similarity, they were asked about their academic qualification. The application stored answers only of users that completed the survey: they had to answers the 5 molecular similarity questions, and the question about academic qualification.

The web application was served on the Heroku Cloud Application Platform [181]. We sent invitations to take part to the survey to 69 chemistry departments and insti-



Do you think that the two molecules are similar?

○ Yes
○ No

There are no right or wrong answers!

2D Representations:

3D Representations (interactive):

Figure 8.1: Screenshot of the initial view of a similarity question in the survey. The 3D representations are interactive, and could be reset to the initial view.

tutions worldwide (Table 8.1). The survey was available on Heroku from 2021-04-14 to 2021-06-28.

The results of the survey were automatically stored by the web application on a private PostgreSQL [182] database available through Heroku. The results were queried, aggregated, and locally stored using a Python script based on SQLAlchemy [183]. After the survey was completed, and the Heroku web application was shut down, we used the new data-set to train the same groups of single-feature and double-feature models that were built using the original training-set.

Institution	Invitation Date
Università degli Studi di Milano-Bicocca	2021-04-14
Politecnico di Milano	2021-04-15
Università degli Studi di Torino	2021-04-26
Università degli Studi di Genova	2021-04-26
Università degli Studi di Pavia	2021-04-14
Università degli Studi di Padova	2021-04-27
Università degli Studi di Udine	2021-04-29
Università degli Studi di Trieste	2021-04-28
Università degli Studi di Ferrara	2021-04-28
Università di Bologna	2021-04-29
Università degli Studi di Modena e Reggio Emilia	2021-05-03
Università degli Studi di Parma	2021-04-29
Università di Pisa	2021-05-05
Università degli Studi di Firenze	2021-05-04
Università degli Studi di Siena	2021-05-04
Università degli Studi di Roma “La Sapienza”	2021-05-11
Università degli Studi di Roma “Tor Vergata”	2021-05-11
Università degli Studi di Perugia	2021-05-03
Università degli Studi dell’Aquila	2021-05-03
Università degli Studi di Napoli “Federico II”	2021-05-10
Università degli Studi di Bari “Aldo Moro”	2021-05-05
Università degli Studi di Sassari	2021-05-04
Università degli Studi di Palermo	2021-05-06
Università degli Studi di Catania	2021-05-05
Università degli Studi di Messina	2021-05-06
Institute of Chemistry and Biochemistry / Berlin	2021-05-13
Department of Chemistry / Hamburg	2021-05-17
Institute of Pharmacy / Berlin	2021-05-13
Jagiellonian University in Kraków	2021-05-17
Universidade Nova de Lisboa	2021-05-17

Institute of Biochemistry / Cologne	2021-05-17
Institute of Organic Chemistry / Cologne	2021-05-17
Institute of Theoretical Chemistry / Cologne	2021-05-17
Chemistry in Pharmaceutical Sciences / Madrid	2021-05-18
Biochemistry and Molecular Biology / Madrid	2021-05-18
Faculty of Chemistry / Barcelona	2021-05-20
Scienze del Farmaco / Milano	2021-05-26
Department of Chemistry / Delhi	2021-06-01
University of Frankfurt	2021-06-03
University of Edinburgh	2021-06-08
University of Toronto	2021-06-09
University of Sydney	2021-06-09
University of Melbourne	2021-06-10
University of Cambridge	2021-06-10
University of Leicester	2021-06-11
University of Oxford	2021-06-14
University of Copenhagen	2021-06-14
KTH Royal Institute of Technology / Stockholm	2021-06-14
University of British Columbia	2021-06-14
Beijing Normal University	2021-06-14
University of Mumbai	2021-06-14
University of Bangalore	2021-06-14
University of Nottingham	2021-06-15
Carnegie Mellon University / Pittsburgh	2021-06-15
University of Tokyo	2021-06-15
University of Hong Kong	2021-06-15
University of Vienna	2021-06-15
Vrije Universiteit Brussel	2021-06-15
University of Oslo	2021-06-16
IIQ / University of Seville	2021-06-16
Seoul National University	2021-06-16
University of Munich	2021-06-16
Zhejiang University	2021-06-16
Harvard University	2021-06-21
Massachusetts Institute of Technology	2021-06-21
University of Berkeley	2021-06-21
Yale University	2021-06-21
Moscow State University	2021-06-21

Saint Petersburg University	2021-06-22

Table 8.1: Chemistry departments and institutions that were invited to participate in the molecular similarity survey.

Chapter 9

Results and discussion

9.1 Performance of single-feature models on the original training-set

In order to easily compare the new 3D similarity-prediction models with the 2D models developed by Franco *et al.*, we recreated the similarity-prediction LogReg models based on open source fingerprints [152] (see subsection 8.3.4 for description of how all models were built). For simplicity, we will focus on the best 2D model: the one based on Tanimoto CDK Extended.

Qualitatively speaking, Tanimoto CDK Extended values are highly correlated with the percentage of human experts that considered a molecule-pair to be similar (Figure 9.1). The probabilities predicted by the Tanimoto CDK Extended LogReg model are visually a good fit for the human assessments of similarity. For a quantitative evaluation of the similarity-prediction model, we measured its performance on the original training-set with a variety of classification metrics (Table 9.1).

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
Tanimoto CDK Extended	93	0.988	0.987	0.133	0.045
VolSurf Similarity	76	0.821	0.808	0.514	0.168
TanimotoCombo	91	0.970	0.963	0.211	0.065

Table 9.1: Performance of single-feature models on the original training-set.

The Tanimoto CDK Extended single-feature model is very successful in the similarity-prediction task. It correctly predicted 93 out of 100 molecule-pairs. The other metrics also present good values. AUROC and AveP are very high, almost their maximum value of 1. L_{Brier} is pretty low, almost 0.

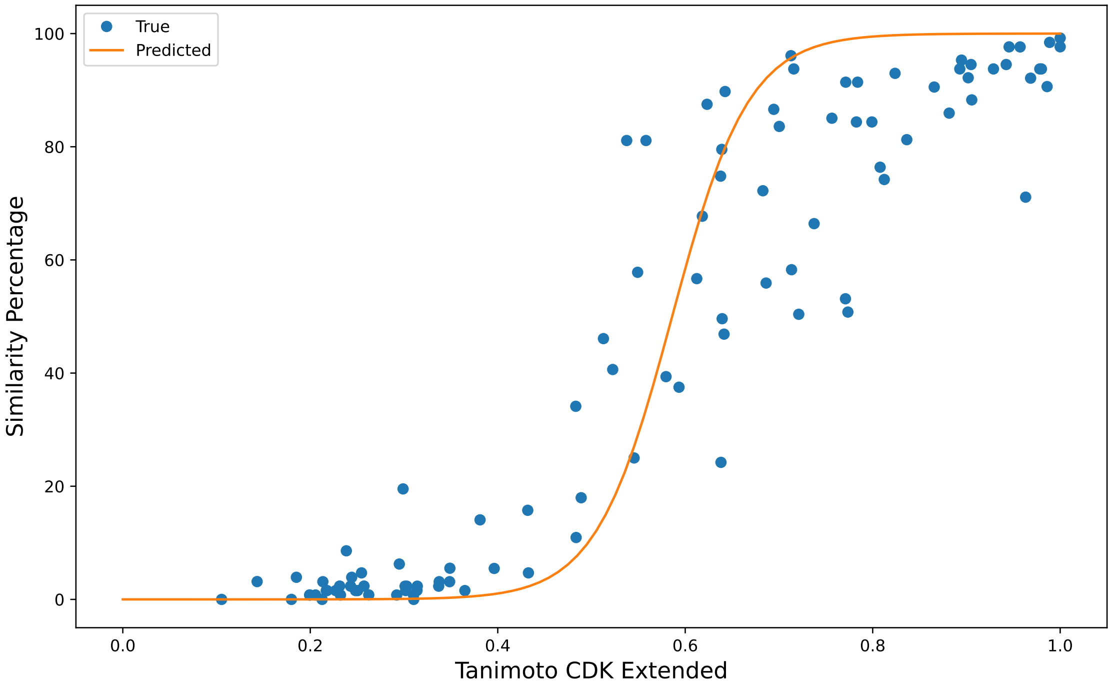


Figure 9.1: LogReg curve of original Tanimoto CDK Extended single-feature model.

The LogReg model based on the VolSurf Similarity calculated with the MOE 3D protocol is shown in Figure 9.2. We can qualitatively observe that the VolSurf Similarity is correlated with the human assessments of similarity: the correlation is not very strong, and many data points are far away from the modeled LogReg curve. We confirm this qualitative observation by considering performance metrics calculated on the original training-set (Table 9.1). The model correctly predicted only 76 molecule-pairs, compared to the 93 correct predictions by the Tanimoto CDK Extended model. The other performance metrics are not bad, but they are definitely worse than the metrics for the Tanimoto CDK Extended model. These results were expected, considering that VolSurf descriptors were not developed to perform similarity calculations. They are an excellent tool for 3D-QSAR modeling, since each VolSurf descriptor is an easily interpretable scalar quantity obtained from the whole MIF tensor. But the calculation of physical properties from the MIFs is conceptually similar to the information compression that caused the original 3D models by Franco *et al.* to perform badly. So, VolSurf descriptors are not ideal for the similarity-prediction task, and will not be considered further.

The LogReg model based on TanimotoCombo values calculated with the OpenEye 3D Protocol is shown in Figure 9.3. TanimotoCombo is highly correlated with human assessments of molecular similarity. Data points are well fitted by the modeled LogReg curve. The quantitative performance metrics are also good (Table 9.1).

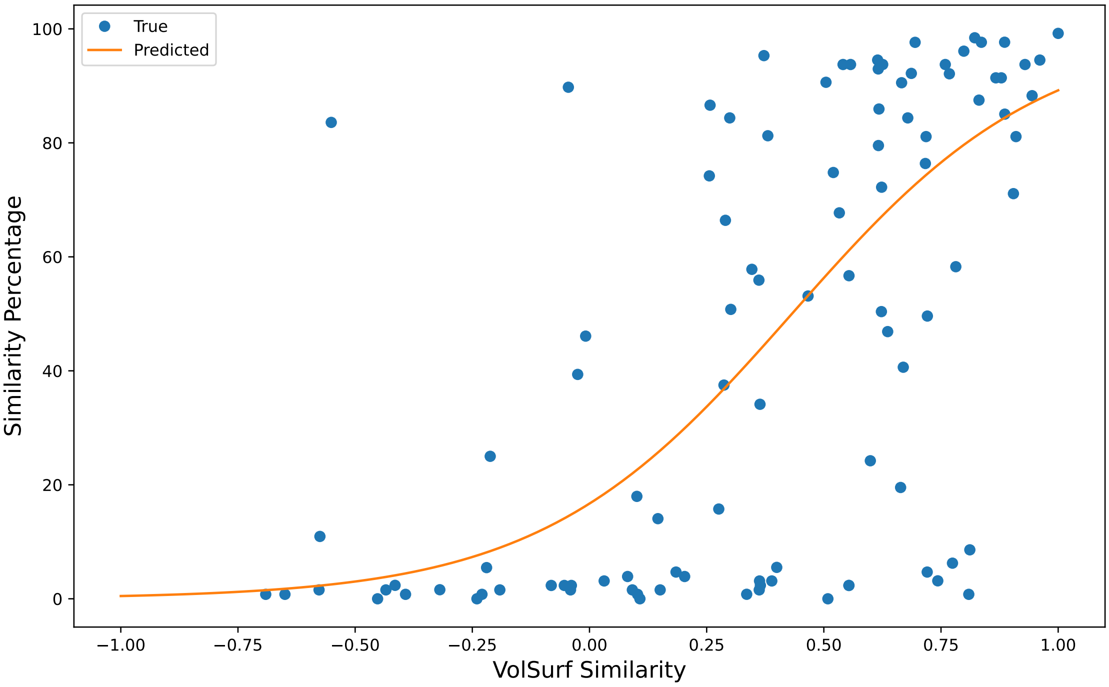


Figure 9.2: LogReg curve of VolSurf Similarity model.

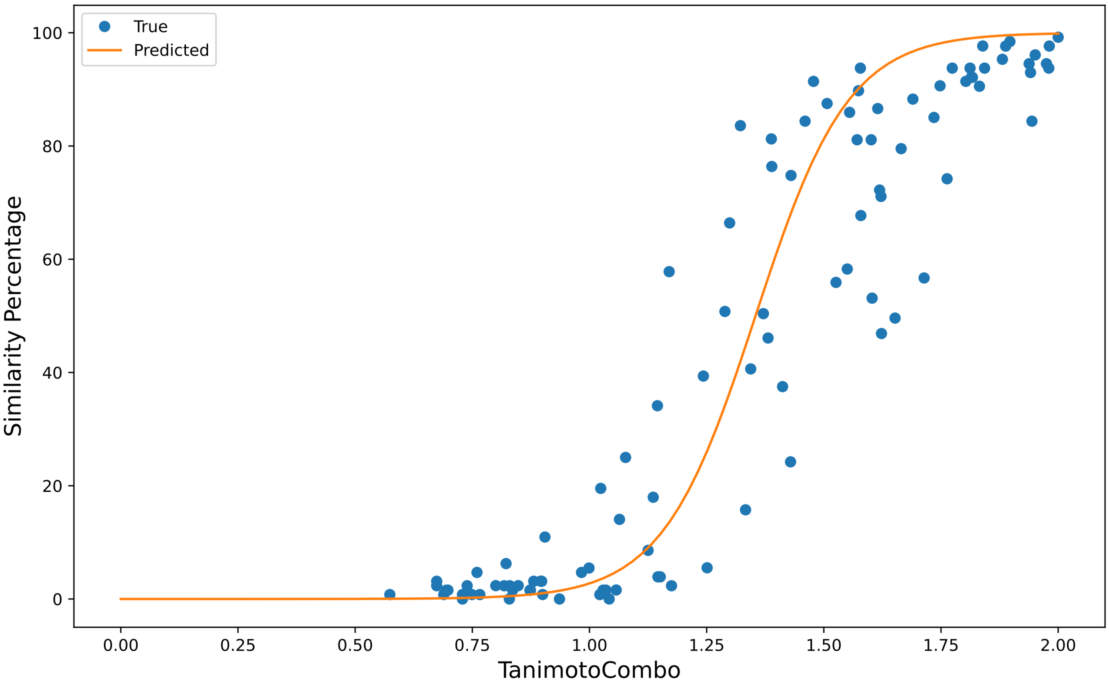


Figure 9.3: LogReg curve of original TanimotoCombo single-feature model.

The TanimotoCombo model correctly predicted 91 molecule-pairs, compared to the 93 correct predictions by the Tanimoto CDK Extended model. The other metrics are also good, comparable to the values obtained by the Tanimoto CDK Extended model (which remains the best single-feature model).

TanimotoCombo is a similarity measure suitable for the similarity-prediction task. It can be considered as a 3D extension of Tanimoto CDK Extended. The CDK Extended fingerprint encode the presence of substructural patterns within a molecule. So, Tanimoto CDK Extended is a measure of the similarity of substructural patterns in a pair of molecules. This feature is similar to the chemistry alignments performed by ROCS, and included in TanimotoCombo values. This information is "translated" from the 2D to the 3D space when transitioning from Tanimoto CDK Extended to TanimotoCombo. TanimotoCombo values also store another crucial information of 3D structures: the molecular shape.

9.2 Comparing 2D and 3D models

The Tanimoto CDK Extended and TanimotoCombo models predicted differently six molecule-pairs, whose ID are 31, 48, 54, 60, 72, and 94 (see the original training-set in Table S1 of [151]). Of the six molecule-pairs that were differently predicted, two were correctly predicted by the TanimotoCombo model (IDs 48 and 60), whereas the other four were correctly predicted by the Tanimoto CDK Extended model.

Figure 9.4 includes the 2D graphs, the OMEGA conformers, and the ROCS colored shape surfaces (generated with vROCS tool) of molecule-pair 60. 81.1% of experts considered molecule-pair 60 to be similar, based on 2D representations. Molecule-pair 60 has a Tanimoto CDK Extended of 0.538: the model calculated a similarity probability of 23.4%, so the pair was classified as dissimilar. On the other hand, the TanimotoCombo value for molecule-pair 60 is 1.601, and the model calculated a similarity probability of 92.3%, thus correctly classifying the pair as similar. Since the molecules in pair 60 are quite small, Tanimoto CDK Extended value is affected by the presence few different groups. But the experts recognized that the two molecules have a similar scaffold and similar features. This similarity was correctly captured by TanimotoCombo: the molecular shapes and the relative positions of chemical features overlap well (Figure 9.4).

Figure 9.5 includes the 2D graphs, the OMEGA conformers, and the ROCS colored shape surfaces of molecule-pair 72. 66.4% of experts considered pair 72 to be similar: the majority considered the pair similar, but it is not a clear-cut decision, since a quite large amount of experts (33.6%) considered the pair dissimilar. The Tanimoto CDK Extended value for pair 72 is 0.738, and the model classified pair 72 as similar with very high probability (97.6%). The TanimotoCombo for pair 72 is 1.299, and with a calculated similarity probability of 36.4%, the pair was classified as dissimilar by the model. Even though the Tanimoto CDK Extended model classified correctly pair 72, the high similarity probability does not capture the ambiguity of the situation. The Tanimoto CDK Extended measure considers the molecules to be similar because they have similar basic functional groups. But the relative positions of such functional groups are not identical, and this is captured by the ROCS colored shape surfaces Figure 9.5.

Figure 9.6 includes the 2D graphs, the OMEGA conformers, and the ROCS colored shape surfaces of molecule-pair 94. Experts classified the molecule-pair as dissimilar, but it was not a clear-cut decision: 46.1% of votes for similarity, and 53.9% of votes for dissimilarity. With a Tanimoto CDK Extended value of 0.513, and a very low calculated similarity probability of 14.3%, the 2D model classified pair 94 as dissimilar. The 3D protocol produced a TanimotoCombo score of 1.381 for pair 94, thus calculating a similarity probability of 56.7%, and classifying the pair as similar. The TanimotoCombo model classified pair 94 incorrectly, but the calculated similarity probability better captures the ambiguity of the situation. The Tanimoto CDK Extended value for pair 94 is quite low, because both molecules have an aromatic ring, with different chemical features. In this instance, ROCS recognized that the two molecules have a quite similar shape, and that some features are in the same relative position in 3D space (Figure 9.6).

We have found that Tanimoto CDK Extended and TanimotoCombo models produce comparable results, but that their predictions are different for some interesting molecule-pairs. The TanimotoCombo model can classify a molecule-pair correctly where the Tanimoto CDK Extended model failed (e.g., molecule-pairs 48 and 60). The TanimotoCombo model can capture the ambiguity of borderline cases (e.g., molecule-pairs 72 and 94).

TanimotoCombo values are a good input feature for the similarity-prediction task. It can be used with other effective 2D input features to improve model performance on the similarity-prediction task.

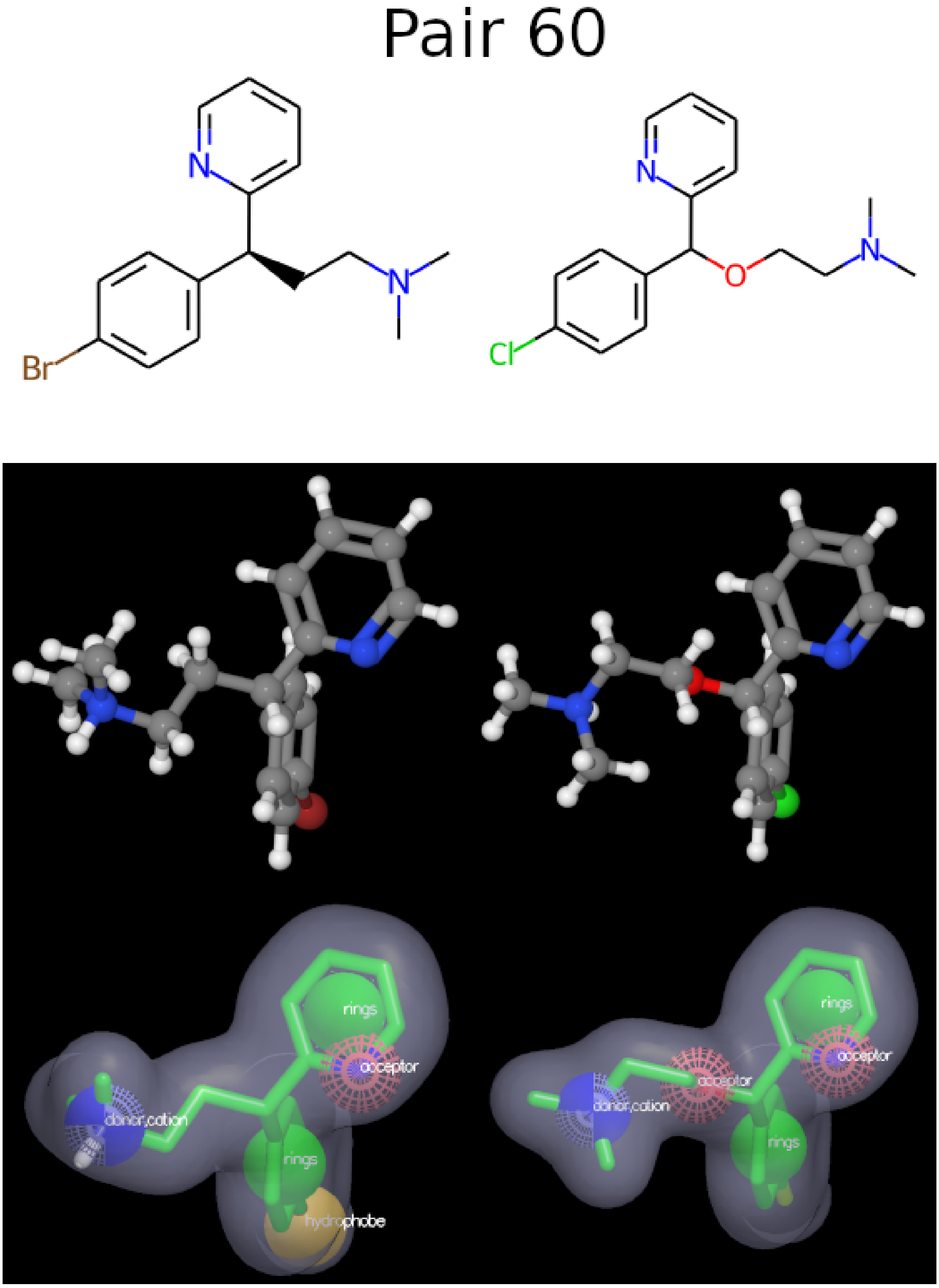


Figure 9.4: 2D graph, 3D conformers, and ROCS surfaces of molecule-pair 60.

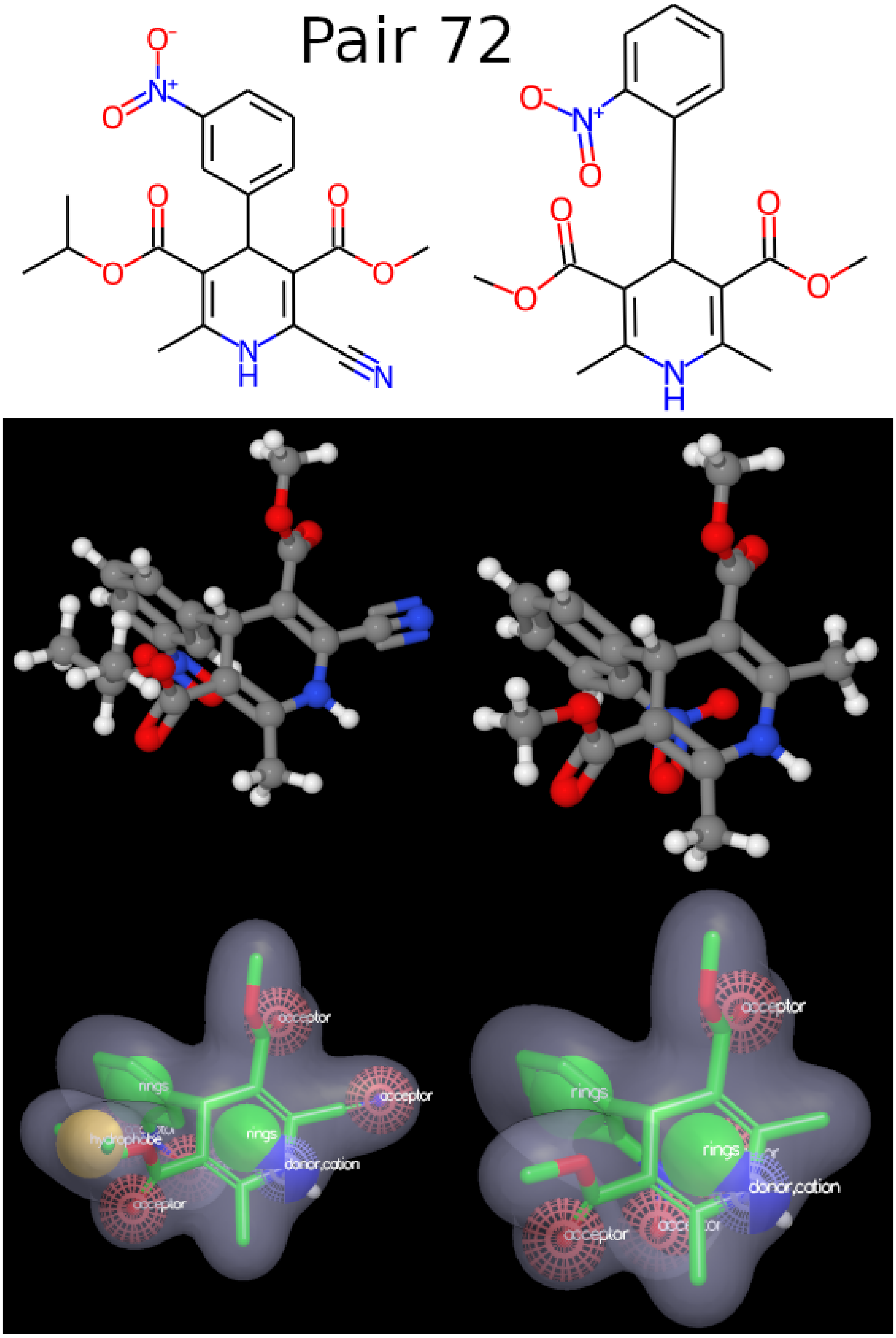


Figure 9.5: 2D graph, 3D conformers, and ROCS surfaces of molecule-pair 72.

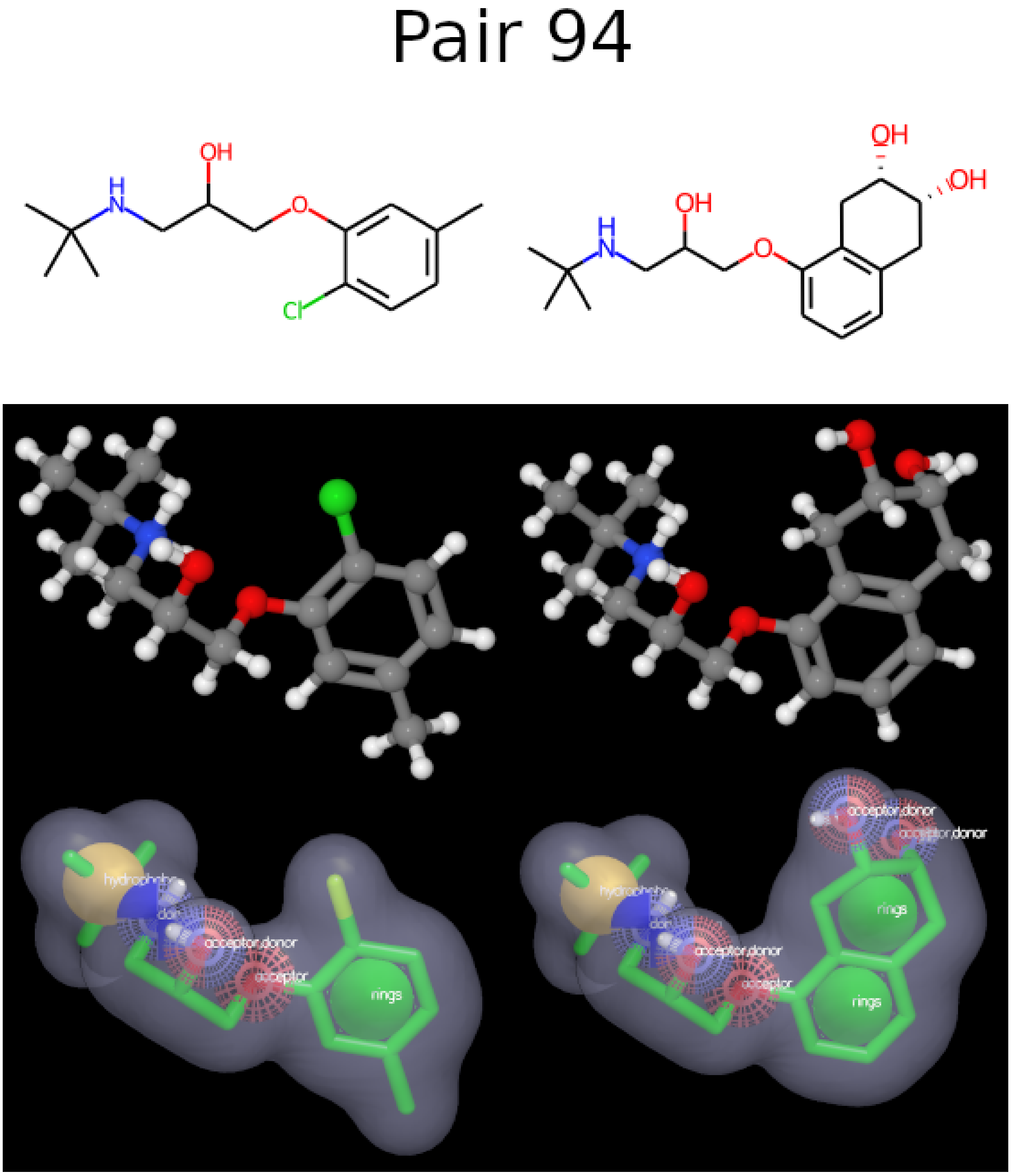


Figure 9.6: 2D graph, 3D conformers, and ROCS surfaces of molecule-pair 94.

9.3 Performance of double-feature models on the original training-set

We built double-feature similarity-prediction to test whether TanimotoCombo values could be used with Tanimoto CDK Extended to improve the performance of similarity-prediction models. The combination of Tanimoto CDK Extended and TanimotoCombo creates a nice separation between the similar and dissimilar molecule-pairs in the original training-set (Figure 9.7). Similarity-prediction models can take advantage of this separation to improve the quality of their predictions.

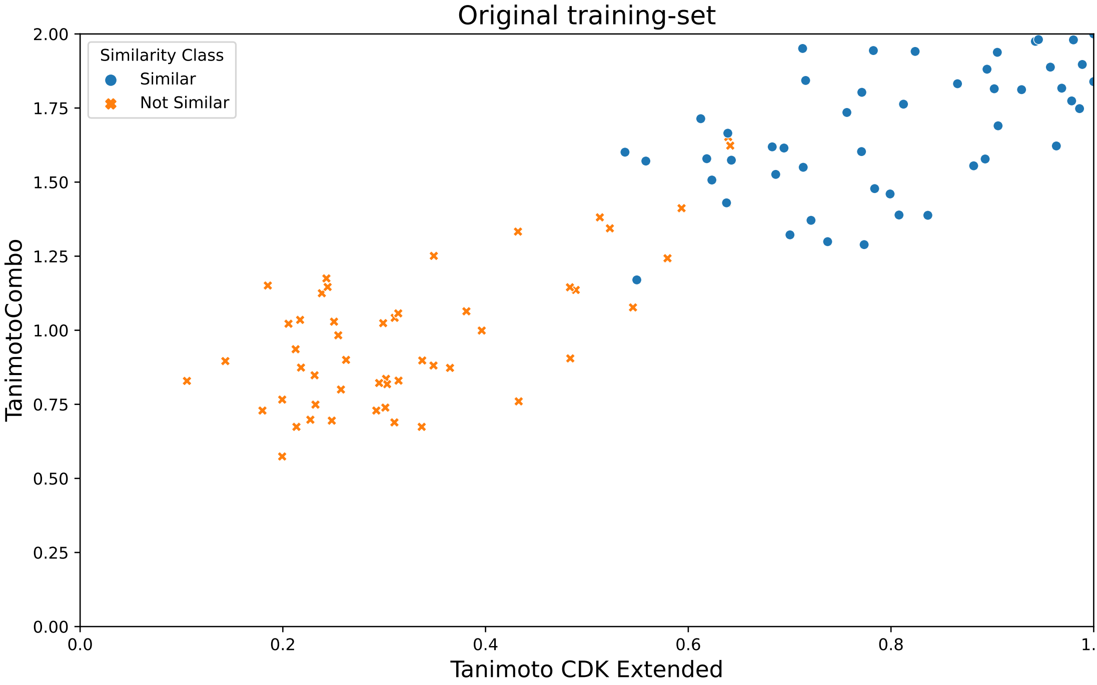


Figure 9.7: The original training-set plotted in 2D/3D similarity space.

Performance metrics of double-feature models trained with default hyperparameters (default double-feature models) are reported in Table 9.2. The double-feature default models include a basic LogReg model with no regularization, LogReg models with L1 and L2 regularization, a DT, and a RF. The basic LogReg model and the L1 and L2 models correctly predicted 95 molecule-pairs, an improvement with respect to the 93 correct predictions of the best single-feature model (the Tanimoto CDK Extended model, Table 9.1). The more advanced performance metrics of these double-feature models are slightly worse than those of the single-feature Tanimoto CDK Extended model: the use of two input features may have caused some overfitting. The best of the three double-feature models based on LogReg is the L1 model: it correctly predicts

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
LogReg	95	0.983	0.980	0.240	0.062
L1	95	0.988	0.986	0.167	0.046
L2	95	0.983	0.980	0.240	0.062
DT	100	1.000	1.000	0.000	0.000
RF	100	1.000	1.000	0.038	0.008

Table 9.2: Performance of default double-feature models on the original training-set.

95 molecule-pairs, and has performance metrics comparable to the single-feature Tanimoto CDK Extended model. The L1 regularization can be an effective technique to improve double-feature similarity-prediction models.

The DT and RF models are apparently much better than the models based on LogReg Table 9.2. They make 100% correct predictions on the original training-set, and their other performance metrics are almost perfect. But it is known that DT models (and to a lesser degree, RF models) are prone to overfitting [139]–[142]. The high degrees of freedom of these models make them able to memorize the whole training-set, thus performing very well on the training-set, but poorly on unseen data.

Hyperparameter grid search with CV is the most common technique to improve model performance on unseen data. We applied the technique to build double-feature models with tuned hyperparameters (tuned double-feature models). The performance of models with best hyperparameters is reported in Table 9.3 (the basic LogReg model with no regularization is not included, since it does not have hyperparameters).

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
L1	95	0.988	0.986	0.179	0.048
L2	95	0.982	0.978	0.452	0.135
DT	99	1.000	1.000	0.014	0.005
RF	95	0.991	0.988	0.132	0.037

Table 9.3: Performance of tuned double-feature models on the original training-set.

The tuned double-feature L1 and L2 models correctly predicted 95 molecule-pairs, just as their default double-feature counterparts (Table 9.2). Tuning the L1 model slightly improved the other performance metrics on the training-set. On the other hand, performance metrics of the L2 model got slightly worse after hyperparameter tuning. A small performance decrease after hyperparameter tuning is normal: a model trades some specific knowledge of the training-set for better predictions on unseen data.

The DT and RF models have worse performance after hyperparameter tuning (Table 9.3). This confirms that default DT and RF models were overfitting on the training-set (Table 9.2). The DT correctly predicts 99 molecule-pairs, and has almost perfect scores for all the other advanced metrics. Even after hyperparameter tuning, the DT is probably overfitting. The only way to know for sure is to test it on unseen data. The RF model has more reasonable performance. It correctly predicts 95 molecule-pairs. The other performance metrics are very good: they are so far the best among all models (except the obviously overfit models). Since RF models have been designed to reduce the overfitting of DTs, it is reasonable to assume that the RF similarity-prediction model that we tuned is not overfit. But to be sure of that, and to confirm that the tuned RF model is the best similarity-prediction model, it is necessary to test it on unseen data.

9.4 Comment on the selection of new molecule-pairs

The 2D/3D similarity landscape after the application of thresholds described in subsection 8.4.1 is presented in Figure 9.8. As could be expected, there is a correlation between Tanimoto CDK Extended and TanimotoCombo. The vast majority of molecule-pairs (8540) are of course dissimilar in 2D and 3D (*dis2D,dis3D*), since we randomly selected the 9000 pairs that were processed with the 2D and 3D protocols, and that are plotted in the figure. There are also many pairs (177) that are similar in 2D and 3D (*sim2D,sim3D*). There are 54 pairs that are similar in 2D and dissimilar in 3D (*sim2D,dis3D*), and 97 pairs that are dissimilar in 2D and similar in 3D (*dis2D,sim3D*). A total of 132 molecule-pairs fell in the buffer zones, and could not be classified as either similar or dissimilar in 2D or 3D: their data points are not shown in Figure 9.8.

100 data points from Figure 9.8 were selected (as described in subsection 8.4.1), and are shown in Figure 9.9. The figure also includes some representative molecule-pairs from the four similarity subsets.

The *dis2D,dis3D* subset includes two main types of molecule-pairs. Some pairs are dissimilar in 2D and in 3D because they are of very different size. The *dis2D,dis3D* subset also includes molecule-pairs that are of comparable sizes, but with different chemical functionalities and shapes. The *dis2D,sim3D* subset includes molecule-pairs with similar size, shape, and relative orientation of functional groups, but with somewhat different chemical functionalities. The *sim2D,dis3D* subset include molecule-pairs whose 2D graph is fairly similar: they are of similar size, and have similar chemical functionalities placed in different positions of the basic scaffold. Their diversity is more apparent when observing their 3D representations. Finally, the *sim2D,sim3D* subset includes molecules that are highly similar in 2D and in 3D: they are of similar size, with similar scaffolds, similar chemical functionalities in similar positions.

The 100 molecule-pairs represented in Figure 9.9 are the ones included in the survey, and subjected to assessments by voluntary users with experience in chemistry (see subsection 8.4.2).

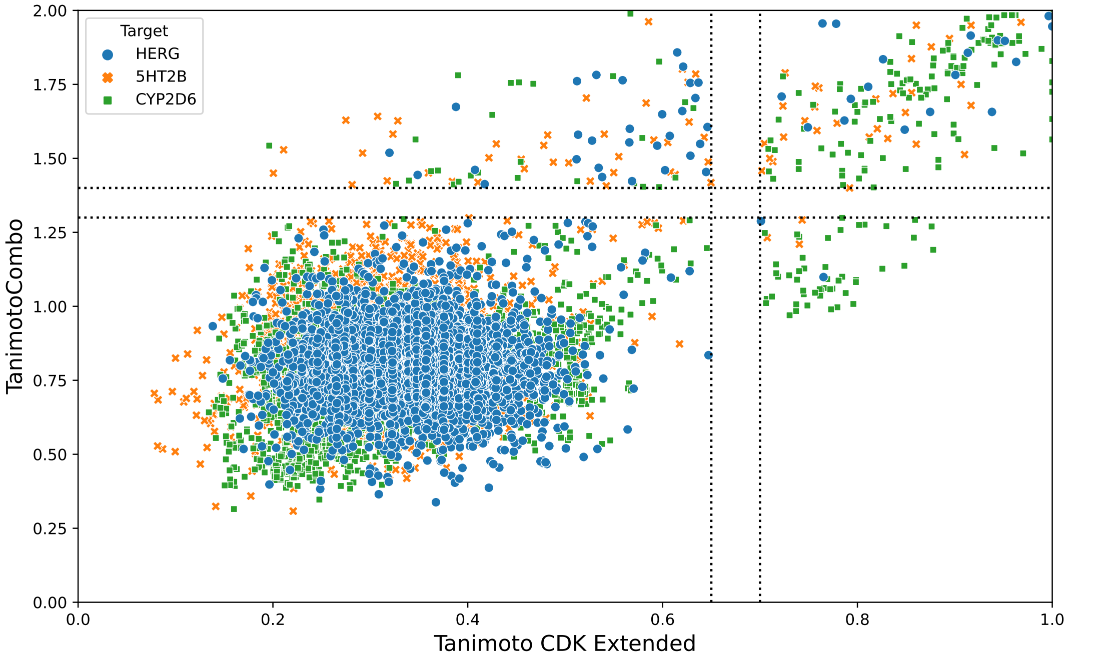


Figure 9.8: Classification of molecules as similar or dissimilar according to Tanimoto CDK Extended and TanimotoCombo values.

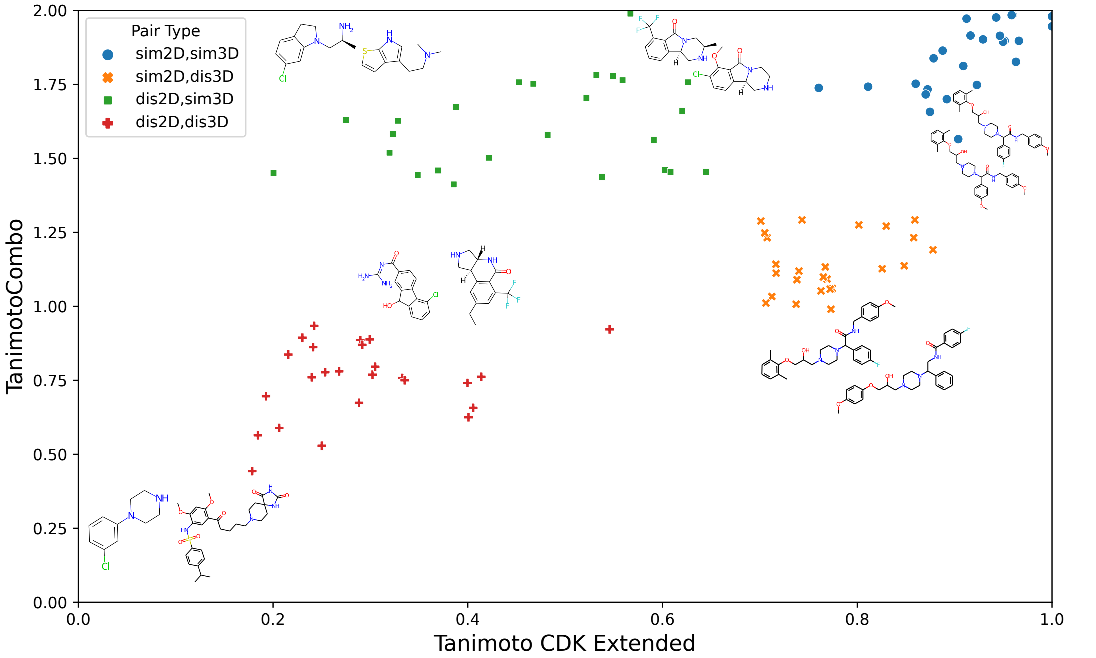


Figure 9.9: 2D/3D similarity landscape of molecule-pairs included in the survey. The 2D graphs of some representative molecule-pairs are included.

9.5 Results of the survey

A total of 518 users clicked the invitation link for the survey, and passed the simple captcha problem that confirmed that they were humans and that they were really willing to participate in the survey (Figure 9.10). 27 users (5.4% of total users) passed the captcha probelm, but did not accept the conditions. 58 users (11.5% of total users) accepted the conditions and started answering the questions, but did not complete the survey. The survey was completed by 418 users (83.1% of total users). Only the answers of users who completed the survey were stored, and were considered for further analysis.

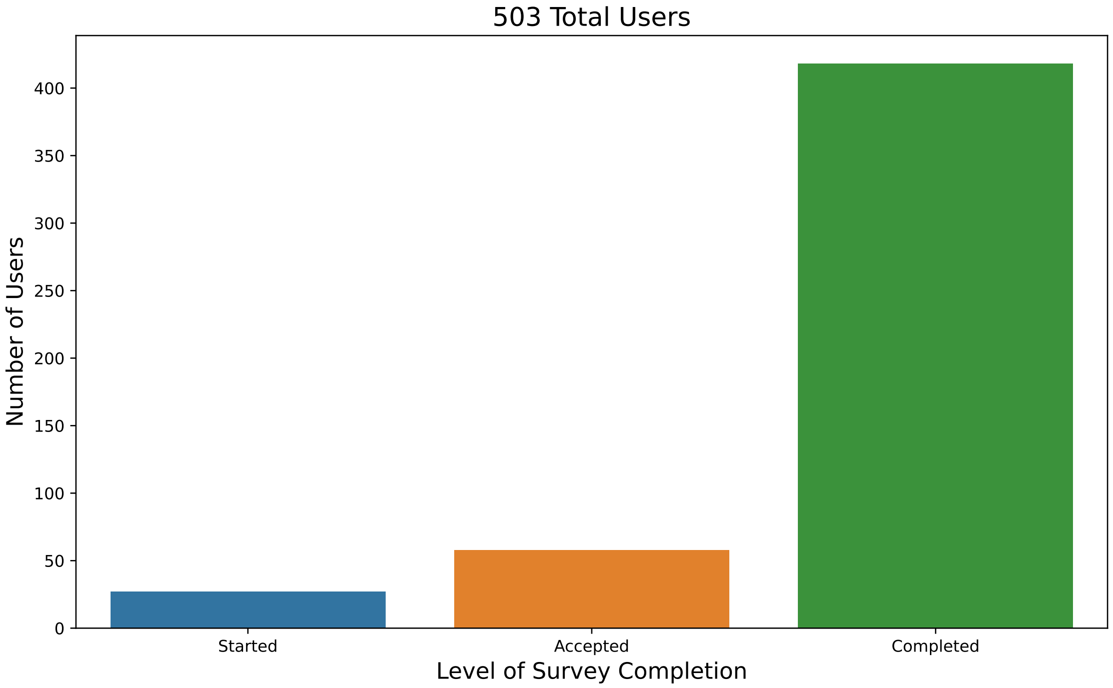


Figure 9.10: Number of users that participated in the survey.

The last question that was answered by the users was about their academic title. Of the 418 users that completed the survey, 70 (16.7%) were PhD Students, 31 (7.4%) were Postdocs, 257 (61.5%) were Professors or Researchers Figure 9.11. The remaining 60 users (14.4%) reported to not possess any of the aforementioned academic titles.

We collected a total of 2090 molecule-pair similarity assessments, since each of the 418 users who completed the survey had to assess the similarity of 5 molecule-pairs. Each user was presented 5 randomly selected molecule-pairs, so each of the 100 molecule-pairs received a different number of assessments (9.12). It is not important that all molecule-pairs are assessed the same number of times: we only needed

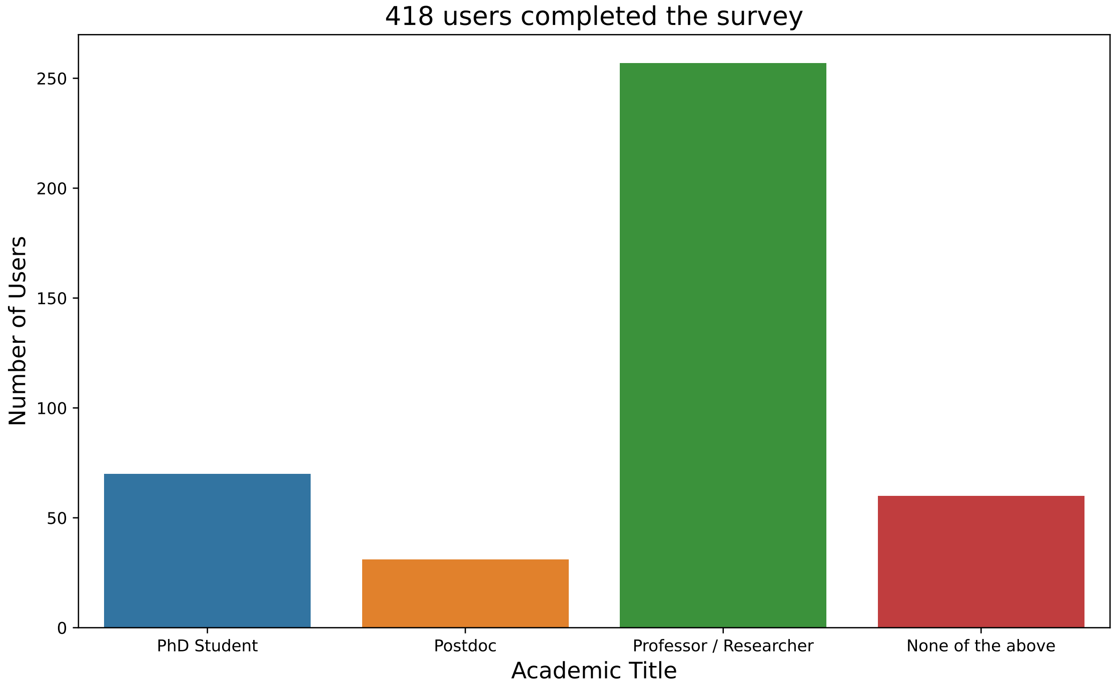


Figure 9.11: Academic title of users who completed the survey.

each molecule-pair to receive a sufficient number of answers so that the similarity assessment was statistically significant. The lowest number of assessments was 11, for molecule-pair 9. Molecule-pairs 19 and 72 received the most assessments (30 assessments each). On average, each molecule-pair received 21 assessments. The number of assessments received by each molecule-pair is reported in Figure B.1.

Another important requirement for the outcome of the survey was that the four similarity subsets received a similar number of answers. Users had to express their judgments on simple similarity scenarios (molecule-pairs in subsets *sim2D,sim3D* and *dis2D,dis3D*) and on more ambiguous situations (molecule-pairs in subsets *sim2D,dis3D* and *dis2D,sim3D*). This requirement was achieved, as shown in Figure 9.13.

The calculated similarity subsets (subsection 8.4.1) are in excellent agreement with the similarity assessments by survey users (Figure 9.14). Molecule-pairs in the *sim2D,sim3D* subset are considered similar by a high percentage of users (81.7% on average). On the other hand, users considered molecule-pairs belonging to the *dis2D,dis3D* subset as dissimilar (92.0% on average). As we expected, users did not agree very strongly on the similarity of molecules in the *sim2D,dis3D* and *dis2D,sim3D* subsets (55.5 and 50.7% respectively).

The same results are shown with greater detail in Figure 9.15, that includes the distributions of assessed similarity percentages in each calculated similarity category (similarity and dissimilarity percentages for each molecule-pair are shown in Figure B.2).

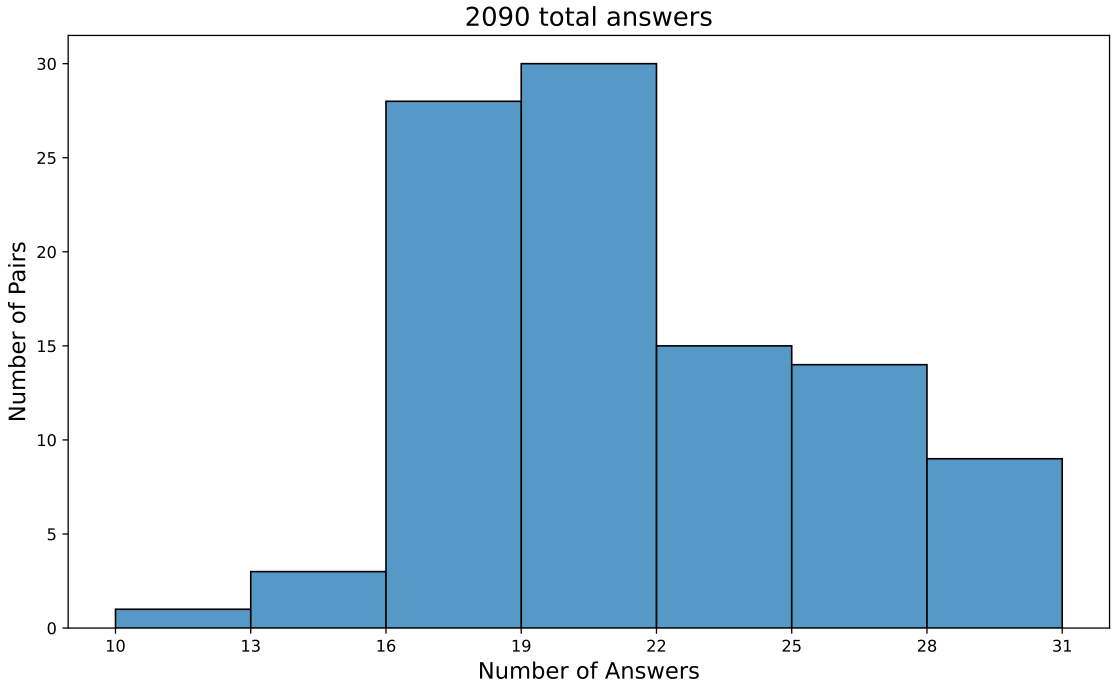


Figure 9.12: Distribution of number of assessments by molecule-pair.

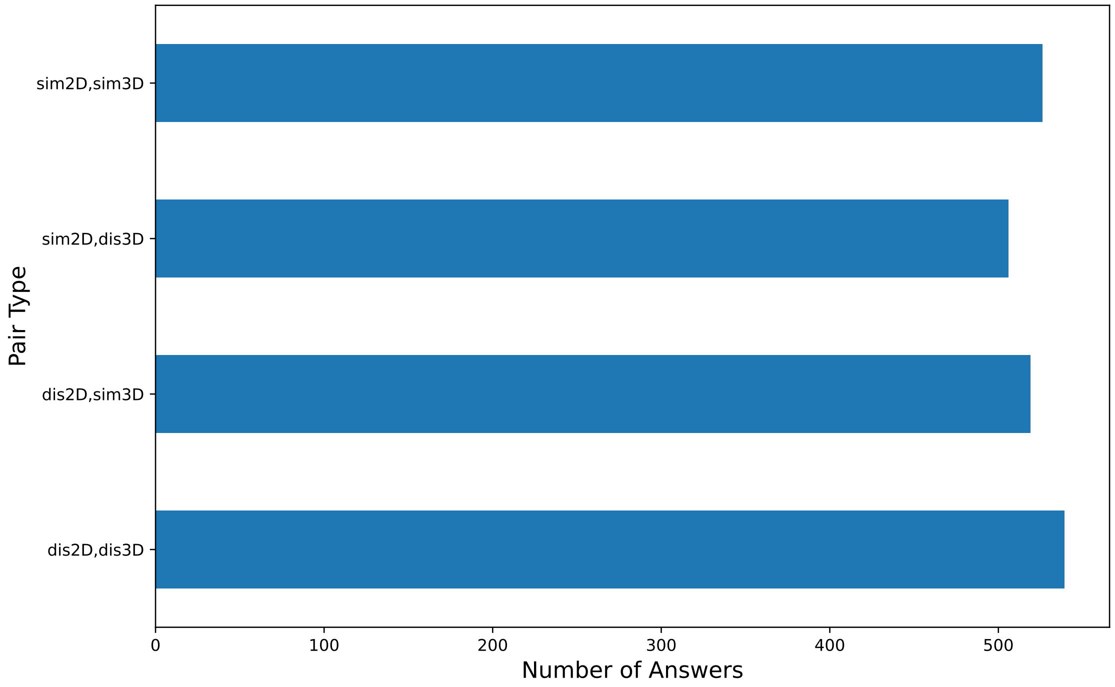


Figure 9.13: Number of assessments for each similarity subset.

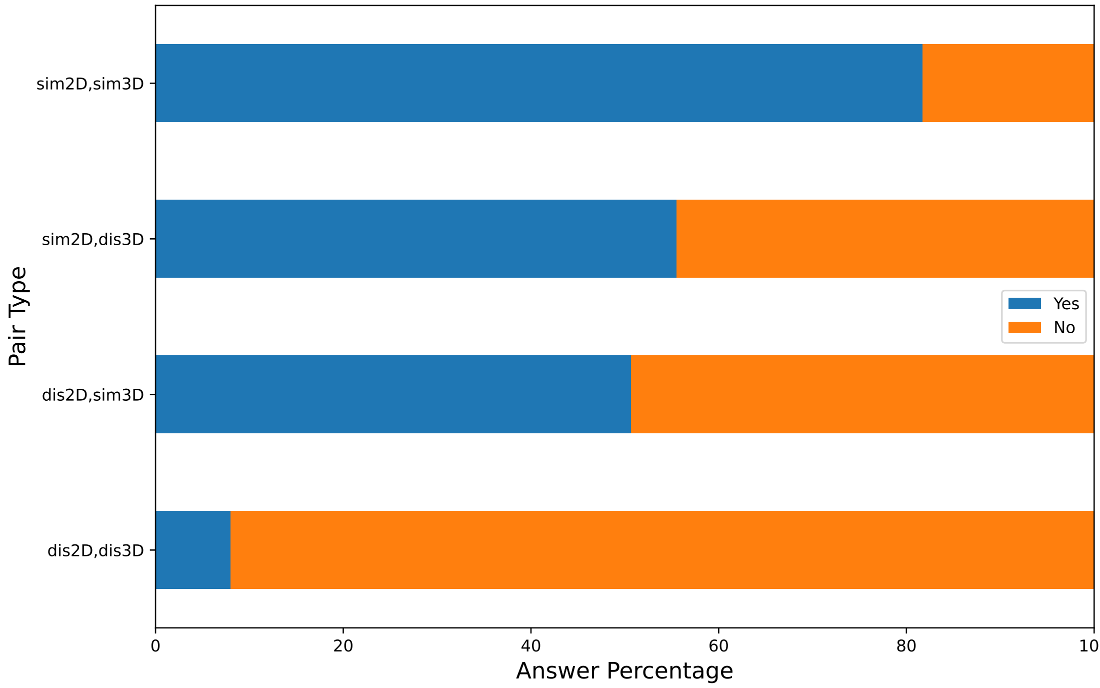


Figure 9.14: Percentage of molecule-pairs considered similar and dissimilar in each calculated similarity subset.

Two molecule-pairs (56 and 61) were judged similar by all the survey users. They belong to the *sim2D,sim3D* group. Eight molecule pairs (76, 81, 85, 86, 87, 93, 95, and 97) were considered dissimilar by all the users. They belong to the *dis2D,dis3D* group. The highest similarity assessment for a molecule-pair in the *dis2D,dis3D* subset was 25.0% for pair 78. The lowest similarity assessment for a molecule-pair in the *sim2D,sim3D* subset was 61.9% for pair 67. The majority of molecule-pairs in the *sim2D,dis3D* and *dis2D,sim3D* subsets obtained similarity assessments between 40 and 60% (Figure 9.15). We consider similarity assessments between 40 and 60% as "uncertain similarity assessments": survey users did not strongly agree on the similarity or dissimilarity of a molecule-pair, if that pair received a similarity assessment between 40 and 60%. 22 molecule-pairs received similarity assessments in the 40–60% range. 13 of them belong to the *sim2D,dis3D* subset, and the remaining 9 to the *dis2D,sim3D* subset.

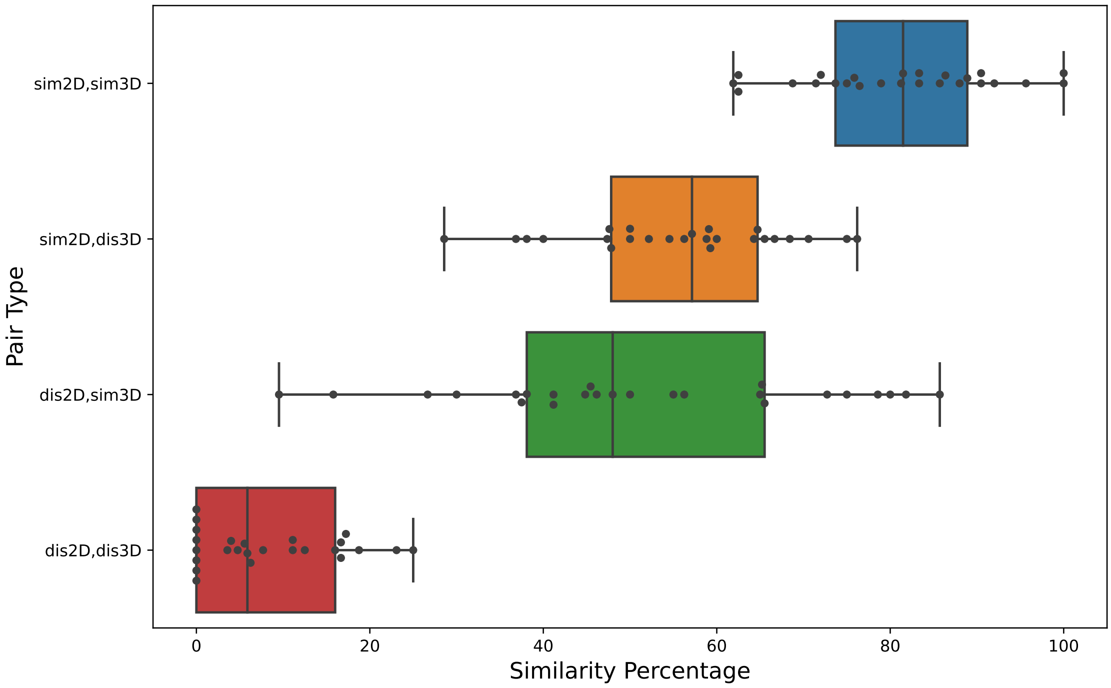


Figure 9.15: Distributions of molecule-pairs considered similar in each calculated similarity subset.

All molecules in the *sim2D,sim3D* subset were considered similar by the majority of users, and all molecules in the *dis2D,dis3D* subset were considered dissimilar by the majority of users. The assessments of molecule-pairs in the *sim2D,dis3D* and *dis2D,sim3D* subsets were more varied. As we mentioned earlier, the majority of molecule-pairs in the *sim2D,sim3D* and *dis2D,sim3D* subsets fell in the "uncertain range" of 40–60% similarity. Similarity assessments in the *sim2D,dis3D* group range from the 28.6% of pair 50, to the 76.2% of pair 31. Similarity assessments in the *dis2D,sim3D* subset are even more varied (biggest box in Figure 9.15): they range from the 9.5% of pair 11 to the 85.7% of pair 10.

When selecting the molecule-pairs to be assessed through the survey by users with experience in chemistry, we wanted to obtain a new data-set that contained different scenarios in molecular similarity. So we assigned the molecule-pairs in four subsets based on the 2D/3D threshold approach described in subsection 8.4.1, and we manually selected 25 pairs from each subset. The survey results matched our expectations. The assessments by survey users are in very good agreement with our classification of molecule-pairs in the four subsets. We obtained a new data-set with many difficult and borderline similarity situations.

9.6 The new training-set

The new data-set was then used to train single-feature and double-feature similarity-prediction models (as described in subsection 8.3.4). We built the same families of models that we built with the original training-set, using the same input features, the same algorithms, and the same hyperparameter tuning procedures. The new data-set contains the percentage of humans that considered each molecule-pair to be similar. This feature was used to classify molecule-pairs as either similar or dissimilar, using majority voting (a 50% similarity threshold, as Franco *et al.* did in their original work [151]). The new training-set is well balanced: it contains 55 similar and 45 dissimilar molecule-pairs. As we expected, the similarity labels are not spread uniformly across the four subsets (Figure 9.16).
All the 25 *dis2D,dis3D* molecule-pairs are classified as dissimilar, and all the 25 *sim2D,sim3D* pairs are classified as similar. The majority of *sim2D,dis3D* molecule-pairs are classified as similar (18 out of 25), with only 7 pairs being classified as dissimilar. On the other hand, molecule-pairs in the *dis2D,sim3D* subset are uniformly distributed across classes: 13 are classified as dissimilar, and 12 as similar.

These preliminary results tell us a lot about the correlation between human similarity assessments and the 2D and 3D similarity measures (Tanimoto CDK Extended and TanimotoCombo). If the similarity measures agree on the similarity or dissimilarity of a molecule-pair, also the humans agree with each other. For more ambiguous molecule-pairs, were the 2D and 3D similarity measures do not agree, human assessments are far less predictable. If the 2D similarity measure is high and the 3D similarity measure is low, humans tend to rely more on the 2D graph representation, and to agree with the 2D similarity measure (that is a numerical extrapolation of the 2D graphs). This is the case for molecule-pairs in the *sim2D,dis3D* subset. On the other hand, if the 2D similarity measure is low and the 3D similarity measure is high, molecule-pairs have an equal chance of being classified as similar or dissimilar by majority voting. This is the case for the *dis2D,sim3D* subset. Some users consider the 3D interactive

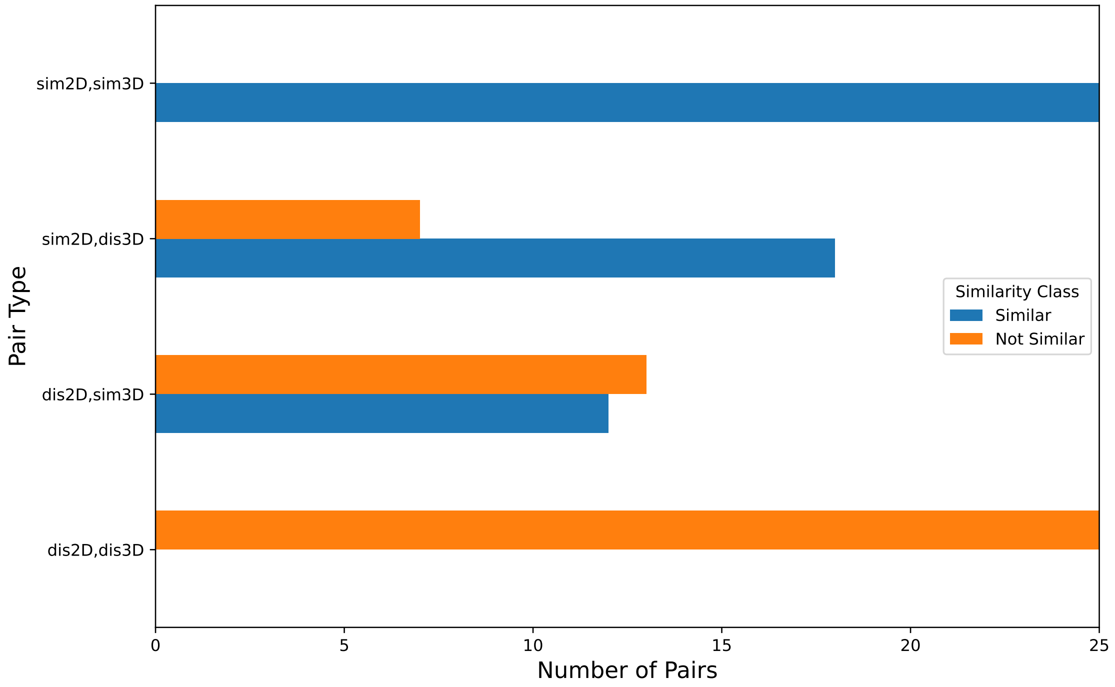


Figure 9.16: Number of molecule-pairs classified as similar or dissimilar for each calculated similarity subset.

representations, and other users only consider the 2D graphs. Most people with experience in chemistry are familiar with 2D molecular graphs: they are a molecular representation that is taught very early in chemistry courses. Even if 3D structures are a more realistic representation of molecules (molecules are 3D objects!), 3D interactive representations are quite obscure for some chemists. We assume that the vast majority of chemists understand what a 3D representation is, and know how to interact with it and how to interpret it. But when given the not-so-common task of assessing molecular similarity, most chemists probably reverted to using the simple 2D graphs that they learned early-on in their careers.

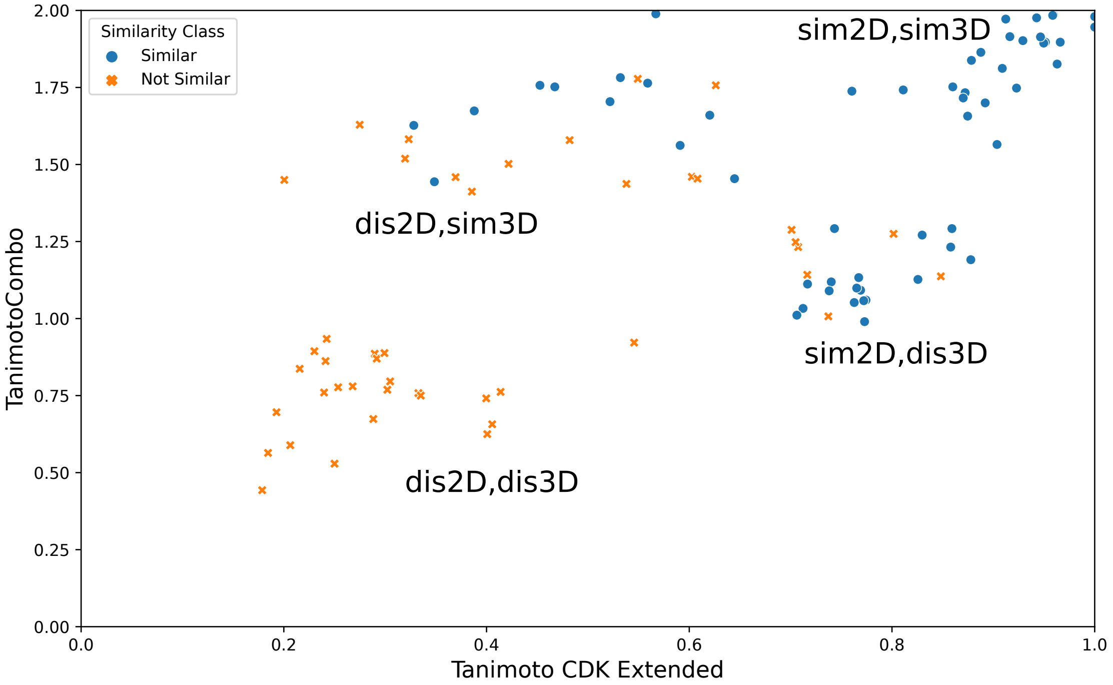


Figure 9.17: The new training-set plotted in 2D/3D similarity space.

Figure 9.17 presents the 2D/3D similarity landscape of the new training-set. Compare it to the landscape of the original training-set (Figure 9.7). Most data points of the original training-set are on the diagonal of Tanimoto CDK Extended - TanimotoCombo plot. The 2D and 3D similarity measures separate well the two similarity classes obtained by majority voting. This explains why models performed well on the similarity-prediction task on the original training-set. On the other hand, only data points of the “simple” *sim2D,sim3D* and *dis2D,dis3D* subsets are on the diagonal of the new similarity landscape (Figure 9.17). The similarity classes of the “borderline” *sim2D,dis3D* and *dis2D,dis3D* subsets are not well separated by the 2D and 3D similarity measures. Solving the similarity-prediction task on the new training-set is clearly much more complex.

9.7 Performance of single-feature models on the new training-set

The LogReg curves of new single-feature and double-feature models qualitatively show that both the Tanimoto CDK Extended and TanimotoCombo values are correlated with the percentage of human experts that considered a molecule-pair to be similar, in the new training-set (Figures 9.18 and 9.19). But, even visually, the performance of the new single-featuremodels on their training-set is worse than that of the original models on their training-set (Figures 9.1 and 9.3). More data points of the new training-set are far from the modeled LogReg curve. This confirms that the new training-set poses a more complex similarity-prediction task than the original training-set.

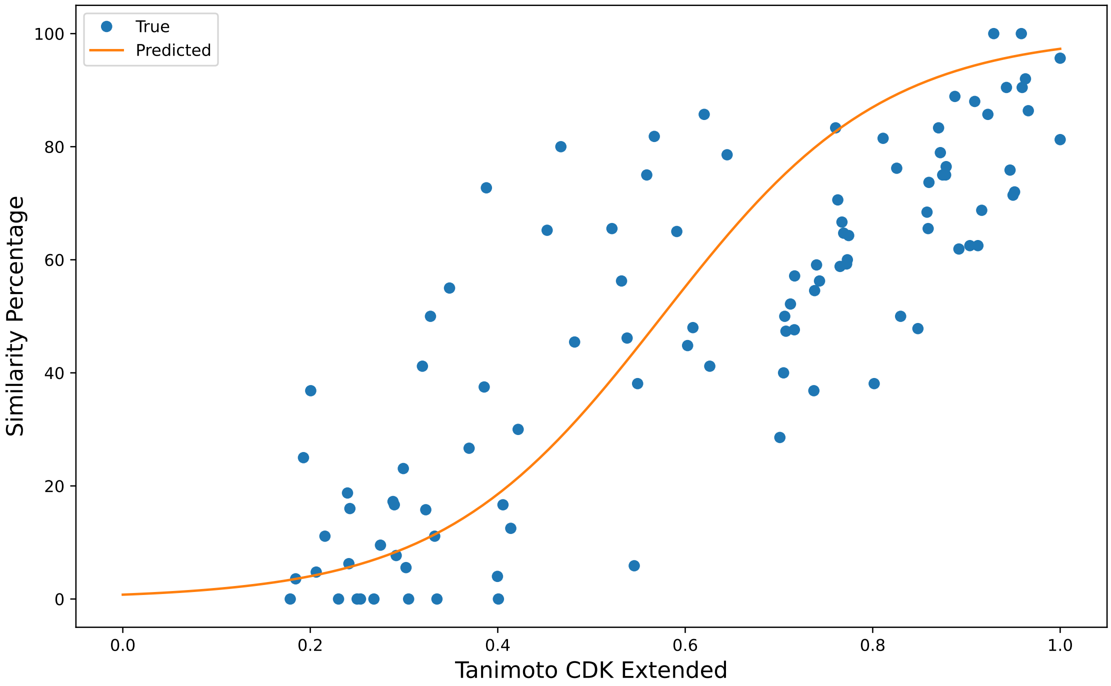


Figure 9.18: LogReg curve of new Tanimoto CDK Extended single-feature model.

Quantitative performance metrics of single-feature models on the new training-set are reported in Table 9.4. Since the new training-set is more complex than the original, the new models have overall worse performance than the ones based on the original training-set (Table 9.1). The original Tanimoto CDK Extended model made 93 correct predictions on its training-set, whereas the new model only made 81 correct predictions. The other more advanced performance metrics are accordingly worse.

The TanimotoCombo model had, on the original training-set, a predictive power comparable to that of the Tanimoto CDK Extended model (Table 9.1). On the other hand,

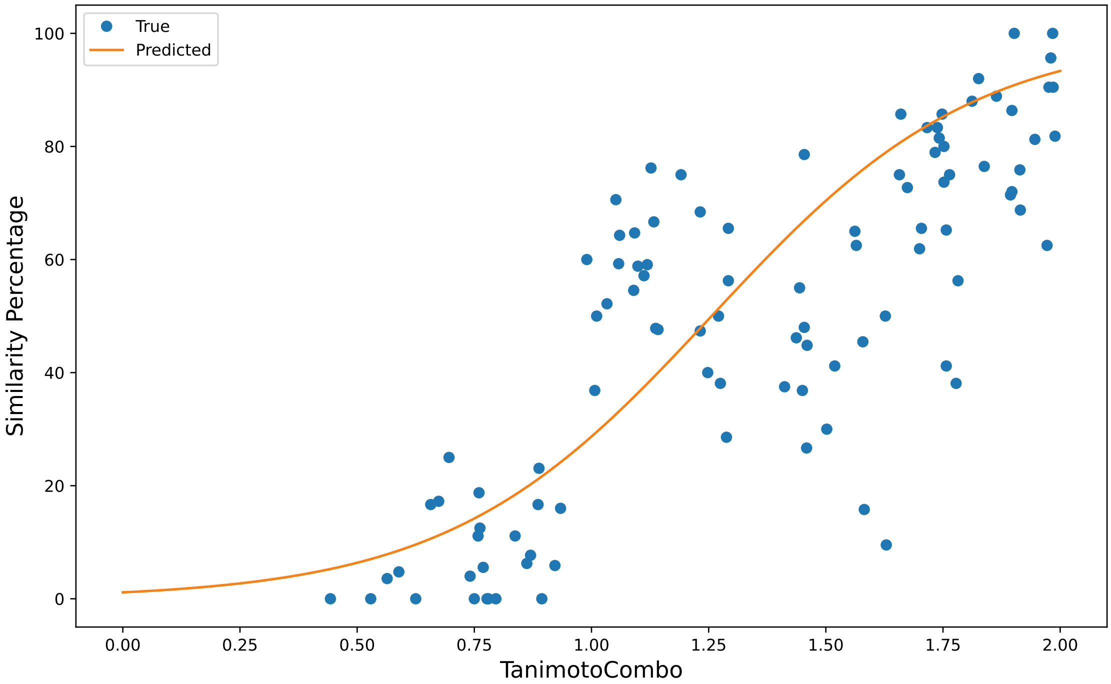


Figure 9.19: LogReg curve of new TanimotoCombo single-feature model.

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
Tanimoto CDK Extended	81	0.920	0.937	0.378	0.122
TanimotoCombo	70	0.845	0.877	0.488	0.167

Table 9.4: Performance of single-feature models on the new training-set.

the new TanimotoCombo model performed definitively worse than the Tanimoto CDK Extended model on the new training-set (Table 9.4). The new TanimotoCombo model only made 70 correct predictions. All the other performance metrics are significantly worse than those of the new Tanimoto CDK Extended model. Everything considered, the performance of the new TanimotoCombo model is not bad per-se. The advanced performance metrics of the new TanimotoCombo model are better than those of the VolSurf Similarity model on the original training-set (Table 9.1). The new TanimotoCombo model has predictive power for the similarity-prediction task. So, we looked further into the reasons behind the poorer performance of the TanimotoCombo model on the new training-set.

9.8 Comparing new 2D and 3D models

We calculated the percentage of correct predictions by the two new single-feature models, for each of the four similarity subsets (Figure 9.20). The Tanimoto CDK Extended

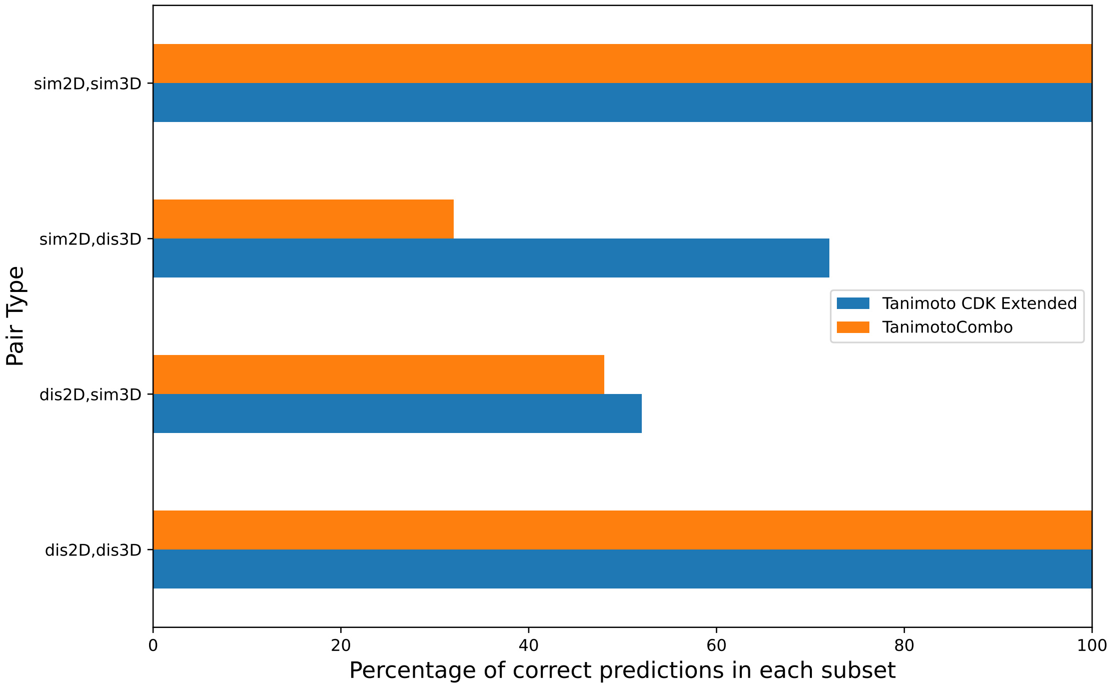


Figure 9.20: Percentage of correct predictions by the single-feature Tanimoto CDK Extended and TanimotoCombo models, for each calculated similarity subset.

and TanimotoCombo models were perfectly able to predict the similarity classes in the *sim2D,sim3D* and *dis2D,dis3D* subsets (100% correct predictions). All the prediction errors occur for molecule-pairs in the *sim2D,dis3D* and *dis2D,sim3D* subsets. On the *dis2D,sim3D* subset, both the Tanimoto CDK Extended and TanimotoCombo models performed poorly: 52 and 48% correct predictions each (almost coin tosses!). On the other hand, the Tanimoto CDK Extended model performed reasonably well on the *sim2D,dis3D* subset (72% correct predictions), whereas the TanimotoCombo model performed very poorly on the same subset (32%). This confirms what we pointed out in section 9.6: when in doubt, humans tend to consider only the 2D molecular graphs, whose similarity is correlated with the Tanimoto CDK Extended values. This explains why Tanimoto CDK Extended values are a better predictor than TanimotoCombo for difficult cases of the similarity-prediction task.

9.9 Performance of double-feature models on the new training-set

Both Tanimoto CDK Extended and TanimotoCombo single-feature models demonstrated predictive power on the new training-set. The combination of Tanimoto CDK Extended and TanimotoCombo features can create better similarity-prediction models. We used the new training-set to build the same types of double-feature models that were successful on the original training-set, and we evaluated their performance. The performance metrics of double-feature models trained with default scikit-learn hyperparameters is reported in Table 9.5.

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
LogReg	83	0.910	0.927	0.390	0.122
L1	84	0.924	0.936	0.335	0.108
L2	83	0.910	0.927	0.390	0.122
DT	100	1.000	1.000	0.000	0.000
RF	100	1.000	1.000	0.086	0.017

Table 9.5: Performance of default double-feature models on the new training-set.

All the default double-feature LogReg models correctly predict more molecule-pairs than the single-feature models (Table 9.4), but are worse than the default double-feature models built and evaluated on the original training-set (Table 9.2). When considering the more advanced performance metrics, only the LogReg model with L1 regularization is better than the best single-feature model (the Tanimoto CDK Extended model). The default double-feature DT and RF models correctly predict 100% of molecule-pairs in the new training-set, and have almost perfect values for the other metrics. This actually means that they are probably overfitting.

To reduce overfitting and improve model performance on unseen data, we applied the same hyperparameter tuning procedure that we used for the original models. Performance metrics of the new tuned double-feature models is reported in Table 9.6.

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
L1	84	0.924	0.936	0.342	0.109
L2	83	0.921	0.934	0.323	0.107
DT	89	0.921	0.925	0.306	0.092
RF	93	0.968	0.971	0.223	0.071

Table 9.6: Performance of tuned double-feature models on the new training-set.

After hyperparameter tuning, the L1 model correctly predicts 84 molecule-pairs, just as the default L1 model. The advanced performance metrics of the tuned L1 model are, on the other hand, slightly worse. A performance decrease after hyperparameter tuning is normal: the tuning procedure is done in order to improve model performance on unseen data. Also the L2 model correctly predicts the same number of molecule-pairs (83) after tuning. But the advanced performance metrics of the L2 model became better after tuning. L2 has better L_{log} and L_{Brier} than L1, but its AUROC and AveP are slightly worse than L1.

After tuning, the DT correctly predicts 89 molecule-pairs: more than the L1 and L2 models. The advanced performance metrics of the DT are comparable to those of L1 and L2: slightly better L_{log} and L_{Brier}, slightly worse AUROC and AveP. The RF model is the best tuned model: it correctly predicts 93 molecule-pairs, and it has the best values for all the advanced performance metrics.

As expected, the DT and RF models became worse after hyperparameter tuning: the default hyperparameters for DT and RF evidently caused overfitting. The only way to understand if the models are stil overfitting the new training-set, is to test them on unseen data. For this reason, we used the cross-test procedure.

9.10 Cross-testing

9.10.1 Original models on new data-set

We employed the cross-test approach (see section 8.2) and used the new data-set to evaluate all models that were built on the original training-set (Table 9.7). N_{correct} is the most important metric when evaluating models on an external test-set: it is the value that we want our models to predict. We also reported all the other advanced performance metrics to have a clearer picture, and to help us in selecting the best models when in doubt.

The single-feature Tanimoto CDK Extended model is one of the best performers. It correctly predicted 81 molecule-pairs (the best $N_{correct}$ is 83). It has the best AUROC and AveP. L_{log} and L_{Brier} for the Tanimoto CDK Extended model are also decent, but they are not among the best values.
The single-feature TanimotoCombo model is the worst overall performer. It correctly predicted the least molecule-pairs among all models (69). The TanimotoCombo model also has poor values for the advanced performance metrics. TanimotoCombo by itself cannot be used to build a model on the original training-set that is successful on the new data-set. When used alongside Tanimoto CDK Extended, TanimotoCombo can improve model performance on unseen data.

		$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
single-feature	Tanimoto CDK Extended	81	0.920	0.937	0.620	0.153
	TanimotoCombo	69	0.845	0.877	0.741	0.235
default double-feature	LogReg	73	0.884	0.912	0.390	0.130
	L1	81	0.916	0.930	0.328	0.111
	L2	73	0.884	0.912	0.390	0.130
	DT	82	0.812	0.776	6.217	0.180
	RF	83	0.889	0.907	0.395	0.130
tuned double-feature	L1	78	0.910	0.926	0.342	0.116
	L2	76	0.877	0.907	0.518	0.168
	DT	81	0.811	0.776	6.224	0.183
	RF	72	0.866	0.877	0.410	0.140

Table 9.7: Models built on original training-set and evaluated on new data-set.

The original default double-feature models are the best, when tested on the new data-set. The default RF model correctly predicted 83 molecule-pairs: it is the best of the original models for the most important metric. Its AUROC and AveP are good, but they are worse than those of Tanimoto CDK Extended and L1 models. Its L_{log} and L_{Brier} are among the best values.
Interestingly, the default DT model performed quite well, even though its training-set performance showed it was overfitting (Table 9.2). The default DT correctly predicts 82 molecule-pairs. Its advanced metrics are quite bad compared to those of the other models, especially L_{log}.
Among the default double-feature LogReg models, L1 is the clear winner. It predicts 81 molecule-pairs correctly. It has the best L_{log} and L_{Brier} values, and its AUROC and AveP values are among the best. On the other hand, the basic LogReg with no regularization, and the L2 model, are bad performers. They only make 73 correct predictions, and their

advanced metrics are worse than those of L1.

The hyperparameter tuning based on cross-validation on the original training-set did not successfully improve model performance on the new data-set. All the tuned double-feature models have worse performance than the same models with default hyperparameters. This means that recurrently splitting the original training-set in a training-set and validation-set, and taking the hyperparameters that performed best on the validation-sets, did not improve ability of the models to generalize outside of the training-set. This is a further confirmation that the original training-set and the new data-set contain very different examples of molecular similarity, and that the new data-set contains more difficult and borderline cases that are not expressed well by models built on the original training-set.

The models built on the original training-set performed on the new data-set significantly worse than the models that were trained and evaluated on the new data-set (Tables 9.4, 9.5, 9.6).

9.10.2 New models on original training-set

We used the original training-set to evaluate all models that were built on the new training-set (Table 9.8). All new models performed well on the original training-set. Both the single-feature models have good performance: they correctly predict 92 molecule-pairs, and have good advanced metrics (advanced metrics of the Tanimoto CDK Extended model are better than those of the TanimotoCombo model).

The combination of Tanimoto CDK Extended and TanimotoCombo values improved performance of default double-feature LogReg models. The best default double-feature model is L1: it makes 95 correct predictions (the highest $N_{correct}$, also achieved by tuned L1 and L2 models). Its advanced metrics are better than those of the single-feature models and of the other default double-feature models. The default L1 model is closely followed by the basic LogReg and by the L2 models, that make 93 correct predictions. As expected, the new DT and RF models performed poorly on the original training-set, since they were overfitting their training data (Table 9.5).

The hyperparameter tuning significantly improved the performance of the L2 model. The tuned L2 model is the overall best model, considering the number of correct prediction (95) and the advanced metrics. Performance of the L1 model did not change significantly after hyperparameter tuning.

The performance of the DT improved significantly after tuning, whereas the performance of the RF slightly decreased. The DT and RF are still the worst performers, even after hyperparameter tuning.

The new models, when evaluated on the original training-set, have shown a performance comparable to that of the models that were trained and evaluated on the original training-set (Tables 9.1, 9.2, 9.3).

The cross-test procedure that we developed was successful in telling which models

		N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
single-feature	Tanimoto CDK Extended	92	0.988	0.987	0.213	0.056
	TanimotoCombo	92	0.970	0.963	0.330	0.092
default double-feature	LogReg	93	0.988	0.986	0.282	0.072
	L1	95	0.988	0.987	0.202	0.052
	L2	93	0.988	0.986	0.282	0.072
	DT	82	0.815	0.780	6.217	0.180
	RF	91	0.966	0.967	0.253	0.077
tuned double-feature	L1	95	0.988	0.987	0.215	0.055
	L2	95	0.988	0.986	0.172	0.046
	DT	86	0.868	0.863	0.369	0.116
	RF	90	0.963	0.961	0.255	0.083

Table 9.8: Models built on new training-set and evaluated on original training-set.

only performed well on their training-set, and which models also performed well on unseen data. Models that were built on the original training-set generally had poorer performance on the new data-set. On the other hand, models built on the new data-set performed well also on the original training-set, with performances comparable to models that were actually built on the original training-set. The combination of Tanimoto CDK Extended and TanimotoCombo input features improved the performance of the new models on unseen data. The tuning procedure significantly improved the new models, whereas models built on the original training-set performed worse after hyperparameter tuning. Among the new models, the simple ones based on LogReg are the best performers in cross-testing: the tuned double-feature L2 model is the overall best. The more complex DT and RF models performed worse than the LogReg models when cross-testing on the original training-set.

These results demonstrate that the new data-set is richer than the original training-set. Models built on the new data-set are able to perform well in more molecular similarity scenarios than models built on the original training-set.

9.11 Prediction of uncertain similarity assessments

So far we have only built binary classification models, that predicted a molecule-pair to be either similar or dissimilar. The target variable that the models were trained to predict was the majority voting on molecular similarity: if more than 50% of experts considered a molecule-pair to be similar, the molecule-pair was labeled as similar, otherwise it was labeled as dissimilar. This is in line with EMA practice: opinions at EMA are accepted by majority.

In the original work by Franco *et al.* [151] it was noted that a high number of molecule-pairs fell in the “gray zone” of molecular-similarity assessments. The majority of experts considered these molecule-pairs to be similar or dissimilar, but it was not a strong majority: there was also a significant amount of experts of the contrary opinion. We choose the new molecule-pairs with that in mind. We wanted a new data-set with molecule-pairs that would receive mixed judgments by the experts. We reached our goal (see section 9.5).

The “trustworthiness” of the predictions of a model can be gauged by considering the calculated probability. All the algorithms that we used to build similarity-prediction models accompany their prediction with a probability. Actually, the models calculate a probability first, then convert the probability to a predicted class using an internal threshold, usually of 50%. For instance, a molecule-pair that is classified as similar can be accepted with confidence if, the calculated probability is 90%, and can be taken with a grain of salt if the calculated probability is only 51%.

Could we build more useful models? Could models be trained to actually recognize molecule-pairs that receive mixed judgments by the experts? To answer this question, we modified the similarity-prediction task from a binary to a ternary classification problem. We labeled molecule-pairs that were judged similar by more than 60% of experts as similar. We labeled molecule-pairs that were judged similar by less than 40% of experts (i.e., judged dissimilar by more than 60% of experts) as dissimilar. We considered the range of 40-60% similarity assessments as the “gray zone”, and labeled all molecule-pairs in that range as uncertain. This is the only change in the training-set sets: the target variable, that now has three instead of two possible outcomes. We used the same molecule-pairs, and the same input features (Tanimoto CDK Extended and TanimotoCombo). We built the same types of classification models that were used for the binary similarity-prediction task to solve the new ternary classification problem.

The binary similarity-prediction task is by itself a difficult classification problem, since human judgments do not correlate with calculated similarity for every single molecule-pair. We expected the new ternary similarity-prediction task to be even more difficult: if it is hard to model human judgment, it is even harder to model human indecision! Also, from a merely technical perspective, three classes are harder to model than just two.

9.11.1 Performance on the original training-set

Figure 9.21 shows the familiar similarity landscape of the original training-set, now colored based on the ternary labels. The majority of molecule-pairs are either labeled as similar (41) or dissimilar (48) based on the aforementioned thresholds. Only 11 pairs fell in the 40-60% “gray zone” and were labeled as uncertain. The binary similarity-prediction problem was well balanced, since around half the molecule-pairs were la-

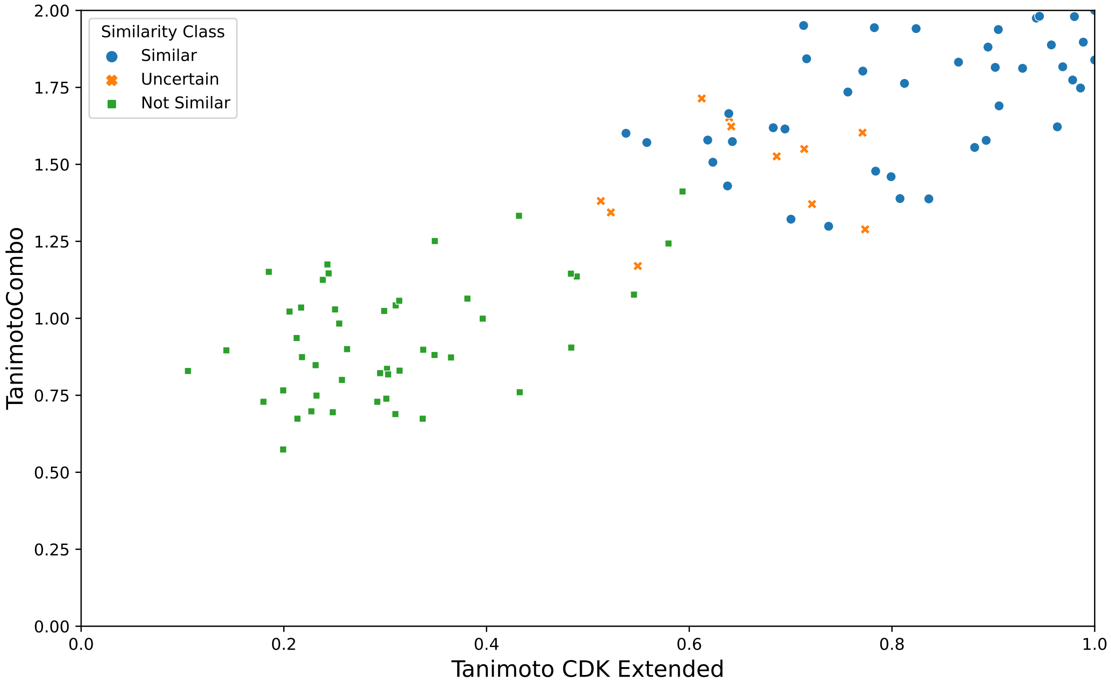


Figure 9.21: The original training-set plotted in 2D/3D similarity space with ternary labels.

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
Tanimoto CDK Extended	82	0.934	0.902	0.407	0.249
TanimotoCombo	77	0.926	0.909	0.433	0.265

Table 9.9: Ternary classification performance of single-feature models on the original training-set.

	Not Similar	Uncertain	Similar
Tanimoto CDK Extended	91.7	81.8	70.7
TanimotoCombo	89.6	54.5	68.3

Table 9.10: Percentages of correct predictions by the ternary classification single-feature models on the original training-set.

beled as similar, and the other half as dissimilar. The new ternary similarity-prediction problem, on the other hand, is unbalanced: there are fewer samples with the uncertain label than with the similar and dissimilar labels. With scikit-learn [71], it was easy to solve this issue, by assigning to each class a weight inversely proportional to the number of samples in that class, during the definition of models. This effectively improves model performance on the uncertain class. The performance metrics of single-feature models built and evaluated on the original training-set are reported in Table 9.9. The Tanimoto CDK Extended model makes 5 correct predictions more than the Tanimoto-Combo model, and also has better values for the other performance metrics.

For the ternary models, other than the usual performance metrics that were used to evaluate the binary models, we calculated the percentages of correct predictions for each class, with respect to the number of samples that actually belong to each class. This information is important, to be sure that the models are actually predicting all the classes, and not just the most frequent classes. In the special classification problem that is the ternary similarity-prediction task, we are particularly interested in models that perform well in predicting the uncertain molecule-pairs. Both the single-feature models make correct predictions across the three classes. They do not only predict the majority classes (similar and dissimilar). Table 9.10. The Tanimoto CDK Extended model predicts the uncertain molecule-pairs with high accuracy (81.8%). On the other hand, the TanimotoCombo model performs well on the similar and dissimilar classes, but it correctly predicts only 54.5% of the uncertain molecule-pairs.

The combination of Tanimoto CDK Extended and TanimotoCombo input features did not improve the default double-feature models based on LogReg (Table 9.11). The basic LogReg, the L1, and the L2 models have worse performance than the Tanimoto CDK Extended single-feature model (Table 9.9). The L1 model is slightly better than the basic LogReg and the L2 models. The more complex DT and RF models are overfitting, as was the case for the binary models.
Only the L1 model correctly predicts uncertain molecule-pairs with high success (81.8%, see Table 9.12). The other models based on LogReg perform reasonably well on the similar and dissimilar classes, but perform very poorly on the uncertain class. The default double-feature DT and RF should not be taken into consideration at this stage, since they are overfitting.

The hyperparameter tuning improved the total number of correct predictions by the L1 and L2 models (Table 9.13). The improvement of $\mathrm{N_{correct}}$ came at the expense of the advanced performance metrics. The fact that hyperparameter tuning improved $\mathrm{N_{correct}}$ while worsening the advanced metrics is easily explained by looking at the percentages of correct predictions for each class (Table 9.14). The improvement in $\mathrm{N_{correct}}$ only occurred for the most frequent classes (similar and dissimilar), whereas the uncertain class is not modeled well.
The DT and RF models are probably still overfitting at this stage, but the only way to

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
LogReg	78	0.935	0.901	0.485	0.269
L1	79	0.942	0.907	0.398	0.232
L2	78	0.935	0.901	0.485	0.269
DT	100	1.000	1.000	0.000	0.000
RF	100	1.000	1.000	0.078	0.031

Table 9.11: Ternary classification performance of default double-feature models on the original training-set.

	Not Similar	Uncertain	Similar
LogReg	91.7	45.5	70.7
L1	91.7	81.8	63.4
L2	91.7	45.5	70.7
DT	100.0	100.0	100.0
RF	100.0	100.0	100.0

Table 9.12: Percentages of correct predictions by the ternary classification default double-feature models on the original training-set.

know for sure is to evaluate them on unseen data.

The ternary similarity-prediction models, in general, performed very poorly on the original training-set (even though they were built on it). The original training-set was not created to be used for this kind of classification task, so these results were expected. The only model that performed well on the ternary similarity-prediction task was the single-feature Tanimoto CDK Extended model. The combination of Tanimoto CDK Extended and TanimotoCombo input features did not improve the performance on the task. The hyperparameter tuning worsened the ability of double-feature models to predict uncertain molecule-pairs.

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
L1	83	0.915	0.890	0.560	0.316
L2	87	0.864	0.882	1.039	0.627
DT	100	1.000	1.000	0.000	0.000
RF	98	0.999	0.995	0.085	0.044

Table 9.13: Ternary classification performance of tuned double-feature models on the original training-set.

	Not Similar	Uncertain	Similar
L1	93.8	45.5	80.5
L2	95.8	0.0	100.0
DT	100.0	100.0	100.0
RF	100.0	90.9	97.6

Table 9.14: Percentages of correct predictions by the ternary classification tuned double-feature models on the original training-set.

9.11.2 Performance on the new training-set

Figure 9.22 shows the 2D/3D similarity landscape of the new data-set, with the new ternary labels. The new data-set contains 35 dissimilar, 43 similar, and 22 uncertain molecule-pairs. The new data-set contains twice as many uncertain molecule-pairs than the original training-set. The classes of the new data-set are more balanced than those of the original training-set. During modeling, we still applied the weighting to further improve balancing.

As expected, the *sim2D,sim3D* and *dis2D,dis3D* subsets contain only similar and dissimilar molecule-pairs. The uncertain molecule-pairs are only in the *sim2D,dis3D* and *dis2D,sim3D* subsets: they contain 13 and 9 uncertain pairs, respectively. The *sim2D,dis3D* subset also contains 9 similar and 3 dissimilar pairs, and the *dis2D,sim3D* subset also contains 9 similar and 7 dissimilar pairs.

As always, we built and evaluated single-feature models on the new training-set (Table 9.15). They do not perform very well on the new training-set, and their performance is definetively worse than that of the single-feature models built and evaluated on the original training-set (Table 9.9). The new ternary single-feature models also fail in predicting the uncertain molecule-pairs (Table 9.16). The Tanimoto CDK Extended and TanimotoCombo models only predict 54.5% and 50% of uncertain molecule pairs, respectively.

	N_{correct}	AUROC	AveP	L_{log}	L_{Brier}
Tanimoto CDK Extended	76	0.886	0.827	0.675	0.372
TanimotoCombo	70	0.863	0.788	0.771	0.419

Table 9.15: Ternary classification performance of single-feature models on the new training-set.

The models based on LogReg benefit by the combination of the Tanimoto CDK Extended and TanimotoCombo input features (Table 9.17). The default double-feature models based on LogReg make more correct predictions, and have better metrics than

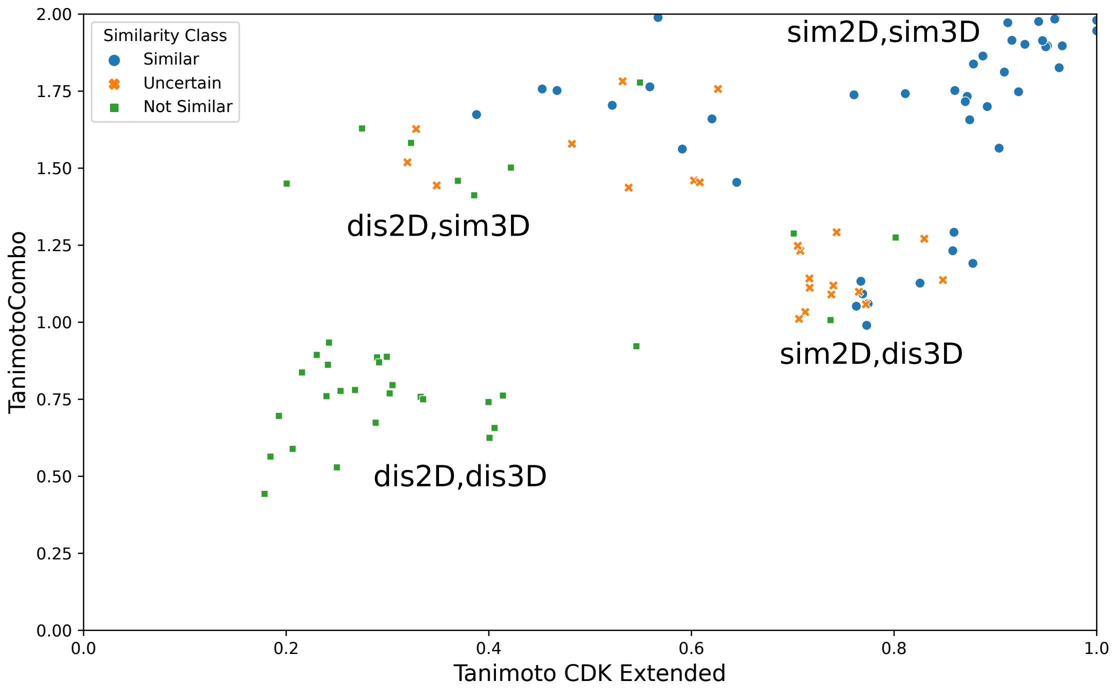


Figure 9.22: The new data-set plotted in 2D/3D similarity space with ternary labels.

	Not Similar	Uncertain	Similar
Tanimoto CDK Extended	85.7	54.5	79.1
TanimotoCombo	74.3	50.0	76.7

Table 9.16: Percentages of correct predictions by the ternary classification single-feature models on the new training-set.

the single-feature models (Table 9.15). Most importantly, these default double-feature models make more correct predictions in the uncertain class (Table 9.18), that is the most interesting class for the ternary similarity-prediction task. On the other hand, the combination of the 2D and 3D input features did not improve the models built and evaluated on the original training-set (Tables 9.11 and 9.12).
As always, the DT and RF models with default hyperparameters are overfitting.

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
LogReg	77	0.913	0.840	0.619	0.344
L1	77	0.917	0.845	0.530	0.299
L2	77	0.913	0.840	0.619	0.344
DT	100	1.000	1.000	0.000	0.000
RF	100	1.000	1.000	0.127	0.051

Table 9.17: Ternary classification performance of default double-feature models on the new training-set.

	Not Similar	Uncertain	Similar
LogReg	77.1	81.8	74.4
L1	85.7	77.3	69.8
L2	77.1	81.8	74.4
DT	100.0	100.0	100.0
RF	100.0	100.0	100.0

Table 9.18: Percentages of correct predictions by the ternary classification default double-feature models on the new training-set.

The hyperparameter tuning slightly made the L1 model do one correct prediction more, at the expense of worse advanced performance metrics (Table 9.19). On the other hand, the L2 model made one correct prediction less, and have performance metrics comparable to those of the default L2 model. The tuned L1 model makes fewer correct predictions on the uncertain class (Table 9.20). The tuned L2 model correctly predicts the same amount of uncertain molecule-pairs. The hyperparameter tuning did not improve the L1 and L2 models on the new training-set. The purpose of hyperparameter tuning is to improve performance on unseen data, so we will evaluate the tuned models on the original training-set using cross-test.
The DT and RF models, after tuning, have very good metrics, and make many correct predictions on the uncertain class. In order to make sure that they are no longer overfitting, we will evaluate them on the original training-set.

	$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
L1	78	0.888	0.821	0.691	0.389
L2	76	0.911	0.843	0.640	0.357
DT	82	0.953	0.894	0.372	0.227
RF	86	0.974	0.956	0.343	0.198

Table 9.19: Ternary classification performance of tuned double-feature models on the new training-set.

	Not Similar	Uncertain	Similar
L1	85.7	63.6	79.1
L2	74.3	81.8	74.4
DT	88.6	86.4	74.4
RF	80.0	81.8	93.0

Table 9.20: Percentages of correct predictions by the ternary classification tuned double-feature models on the new training-set.

9.11.3 Cross-testing

Original models on the new data-set

The performance metrics of models that were built on the original training-set and evaluated on the new data-set are reported in Table 9.21. In general, models that were built on the original training-set do not perform very well on the new data-set. During cross-test, these models make less correct predictions than the models that were built on the new training-set (Tables 9.15, 9.17, and 9.19). Also the advanced metrics are worse. The hyperparameter tuning did not improve the performance of double-feature models on the new data-set. The tuned L1 and L2 are worse than their counterparts with default hyperparameters. The tuning process did not prevent the DT and RF models from overfitting.

The only models that make good predictions on the uncertain class are the default double-feature models based on LogReg (Table 9.22). They correctly predict 81.8% of uncertain molecule-pairs in the new data-set. single-feature models built on the original training-set correctly predict around half of the uncertain molecule-pairs of the new data-set. Tuned double-feature models make even less correct predictions on the uncertain class.

		$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
single-feature	Tanimoto CDK Extended	75	0.882	0.818	0.850	0.377
	TanimotoCombo	67	0.832	0.772	1.200	0.516
default double-feature	LogReg	73	0.886	0.818	0.648	0.371
	L1	70	0.886	0.812	0.617	0.359
	L2	73	0.886	0.818	0.648	0.371
	DT	67	0.703	0.569	11.398	0.660
	RF	65	0.887	0.813	0.693	0.409
tuned double-feature	L1	66	0.809	0.761	0.796	0.464
	L2	67	0.777	0.752	1.053	0.636
	DT	67	0.703	0.569	11.398	0.660
	RF	71	0.811	0.735	5.781	0.427

Table 9.21: Ternary models built on original training-set and evaluated on new data-set.

		Not Similar	Uncertain	Similar
single-feature	Tanimoto CDK Extended	85.7	50.0	79.1
	TanimotoCombo	74.3	45.5	72.1
default double-feature	LogReg	77.1	81.8	65.1
	L1	74.3	81.8	60.5
	L2	77.1	81.8	65.1
	DT	88.6	9.1	79.1
	RF	91.4	0.0	76.7
tuned double-feature	L1	74.3	31.8	76.7
	L2	85.7	0.0	86.0
	DT	88.6	9.1	79.1
	RF	88.6	4.5	90.7

Table 9.22: Percentages of correct predictions by ternary models built on original training-set and evaluated on new data-set.

New models on the original training-set

The models built on the new training-set performed well on the original training-set. (Table 9.23). Their performance metrics (N_{correct} and the advanced metrics) are comparable to those of the models that were built and evaluated on the original training-set (Tables 9.9, 9.11, and 9.13). The tuning process improved the N_{correct} of the new double-feature models on the original training-set. On the other hand, the tuning process did not improve the cross-test performance of models that were built on the original training-set (Table 9.21). When considering N_{correct} and the advanced metrics, the models based on LogReg are better than the DT and RF models.

When considering the correct predictions on the uncertain class, the best models are the single-feature Tanimoto CDK Extended model, and the tuned DT model (Table 9.24). They correctly predict 81.8% of uncertain molecule-pairs in the original training-set. The double-feature models based on LogReg had better performance metrics than these two models, but only predict 45.5% molecule-pairs of the uncertain class, that is the most interesting class in ternary similarity-prediction problems. The double-feature models based on LogReg are better at predicting similar and dissimilar molecule-pairs. A good compromise between overall performance and ability to predict the uncertain class is the tuned RF model. It correctly predicts 63.6% of uncertain molecule-pairs. The single-feature Tanimoto CDK Extended model is another solid performer in this regard.

The cross-testing of ternary similarity-prediction models confirmed that the new data-set is richer than the original training-set, and that it can be used to train models that perform well on unseen data. As we pointed out earlier, the ternary similarity-prediction task is inherently very difficult, and the performance of ternary models is, on average, worse than that of binary models. Using the new training-set, we were able to build good models for the ternary similarity-prediction task. These had good performance and were able to predict uncertain molecule-pairs with high accuracy. We also found out that the two characteristics are not necessarily correlated: models that have very good performance metrics may be bad at predicting the uncertain class, and vice versa.

		$N_{correct}$	AUROC	AveP	L_{log}	L_{Brier}
single-feature	Tanimoto CDK Extended	78	0.934	0.907	0.485	0.277
	TanimotoCombo	76	0.908	0.893	0.577	0.332
default double-feature	LogReg	87	0.936	0.914	0.497	0.268
	L1	85	0.937	0.905	0.390	0.216
	L2	87	0.936	0.914	0.497	0.268
	DT	66	0.719	0.641	11.743	0.680
	RF	80	0.897	0.881	0.467	0.277
tuned double-feature	L1	86	0.929	0.906	0.582	0.316
	L2	88	0.934	0.912	0.523	0.283
	DT	72	0.873	0.848	1.371	0.446
	RF	79	0.911	0.894	0.457	0.274

Table 9.23: Ternary models built on new training-set and evaluated on original training-set.

		Not Similar	Uncertain	Similar
single-feature	Tanimoto CDK Extended	85.4	81.8	68.3
	TanimotoCombo	77.1	45.5	82.9
default double-feature	LogReg	93.8	45.5	90.2
	L1	93.8	45.5	85.4
	L2	93.8	45.5	90.2
	DT	79.2	45.5	56.1
	RF	85.4	36.4	85.4
tuned double-feature	L1	93.8	45.5	87.8
	L2	93.8	45.5	92.7
	DT	87.5	81.8	51.2
	RF	89.6	63.6	70.7

Table 9.24: Percentages of correct predictions by ternary models built on new training-set and evaluated on original training-set.

Chapter 10

Conclusion

We embarked upon our journey by asking ourselves some questions about molecular similarity. How do humans perform molecular similarity assessments? Do chemists consider that molecules are 3D objects, or do they only observe the 2D graphs that they are presented? If given realistic 3D representations, will chemists consider it, or will they still rely on the 2D graphs? Can 3D measures of molecular similarity be useful in modeling majority voting of molecular similarity? And what is the best way to find molecule-pairs that would be hard to judge by experts?

We started from the data and models by Franco *et al.* We built improved similarity-prediction models by combining the best 2D fingerprint with a 3D similarity measure that has never been used for this kind of task: TanimotoCombo. We then explored the 2D/3D similarity landscape in the biggest open database of bioactive compounds, ChEMBL. We developed a new protocol to select sets of molecule-pairs with shared molecular similarity attributes. Most importantly, we wanted to find molecule-pairs that would not be easily classified as similar or dissimilar by expert evaluators. We then used open source Python tools to develop a web survey on 2D and 3D molecular similarity. The molecule-pairs that we selected were assessed by chemists in 69 universities worldwide. It was the first time that a survey on 2D and 3D molecular similarity was ever made. The survey results confirmed our expectations about what the molecular similarity assessments would be, based on the calculated similarity measures. We then used the new data-set to build the same similarity-prediction models that were successful on the original training-set by Franco *et al.*. As we expected, we found out that the new data-set contained more difficult and borderline molecular-similarity cases than the original training-set. The new data-set was a very difficult similarity-prediction modeling task. We used a new validation procedure that used a data-set to evaluate models that were built on another data-set, thus assessing the relative quality of the two data-sets to build similarity-prediction models, and the extent of similarity patterns that they contained. We called this procedure cross-testing. Through cross-testing, we confirmed that the new data-set is richer than the original

one, and that models trained on the new data-set perform better on unseen data. The new data-set is better for model generalizability.
By evaluating our models, we confirmed that the LogReg algorithm (and its variants for multiple input features variants) is very effective for the similarity-prediction task. LogReg even surpasses the more complex DT and RF models. DT and RF had never been used for similarity-prediction, and we found out that they are prone to overfitting, and they usually fail on data other than the training-set. A good hyperparameter tuning procedure is crucial to ensure that models perform well on unseen data.

So far, we had only considered the binary similarity-prediction task: molecule-pairs could only be labeled as similar or dissimilar based on majority voting, and the models would only predict these two possibilities. But in the course of our work, we had found many molecule-pairs that did not receive clear-cut similarity assessments by the majority of experts. We wanted to model human indecision about molecular similarity. So, we added a new possible label to molecule-pairs: "uncertain", and we used the original training-set and the new data-set to build models that solved this ternary similarity-prediction task. It was the first time that this kind of molecular similarity problem was addressed. Even though the ternary similarity-prediction task is more difficult than the classical binary task, we were able to build some successful models. We also found out that the performance metrics that are useful to assess the quality of binary similarity-prediction models do not necessarily correlate with the ability to predict uncertain molecule-pairs in the ternary task. The ternary similarity-prediction modeling process is still at its dawn, and many improvements can be made.

All models (binary and ternary) can benefit from hyperparameter tuning. We used the most common algorithm to find the optimal hyperparameters of a model: grid-search. More advanced approaches can be used: from random search in the hyperparameter space [184], to Bayesian optimization procedures [185]. We used the simple 10-fold cross-validation to evaluate the hyperparameters, but other schemes can be devised. For instance, the ternary similarity-prediction models could benefit from validation-sets with an higher presence of the uncertain class.
All machine-learning models benefit from more data, and most importantly, from more diverse data. The more complex models especially benefit from bigger data-sets. The pair selection procedure that we devised was successful in obtaining interesting molecule-pairs that represented different molecular similarity scenarios. The web survey was an effective approach to obtain molecular similarity assessments on a big scale. The pair selection procedure and the web survey approach can be used to obtain more molecular similarity data, to build better models, and to achieve an ever deeper understanding of the human rational and irrational molecular similarity evaluation process.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in the previous chapters.

Acknowledgments

I acknowledge OpenEye Scientific for allowing me to use their software tools under an academic license. In particular OMEGA and ROCS have been invaluable for my project.
I wish to thank Prof. Alexandre Varnek, Prof. Gilles Marcou, and all the members of the Laboratoire de Chémoinformatique — Institut Le Bel, Université de Strasbourg. Their help and support have been incredible.
Last but not least, I wish to thank all the open source developers that have helped me troubleshoot the problems that I encountered during the project. In particular, I sincerely thank Hai Nguyen for answering many questions about special use-cases of NGLview.

Part III

Appendices

Appendix A

Antifreeze peptides supplements

The following materials are taken from the Supplementary Information of:
E. Gandini, M. Sironi, and S. Pieraccini, “Modelling of short synthetic antifreeze peptides: Insights into ice-pinning mechanism,” *Journal of Molecular Graphics and Modelling*, vol. 100, p. 107 680, Nov. 2020, ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2020.107680

A.1 Water box simulations

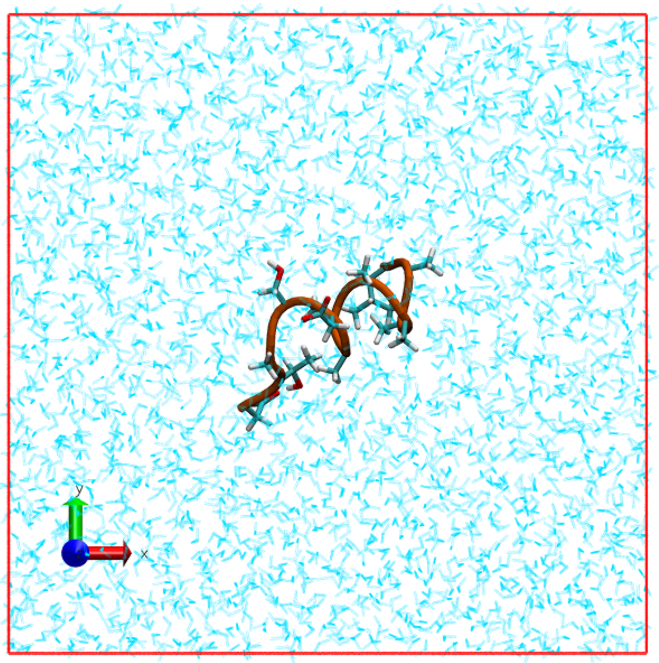


Figure A.1: Final frame of Peptide 1–1 water box simulation.

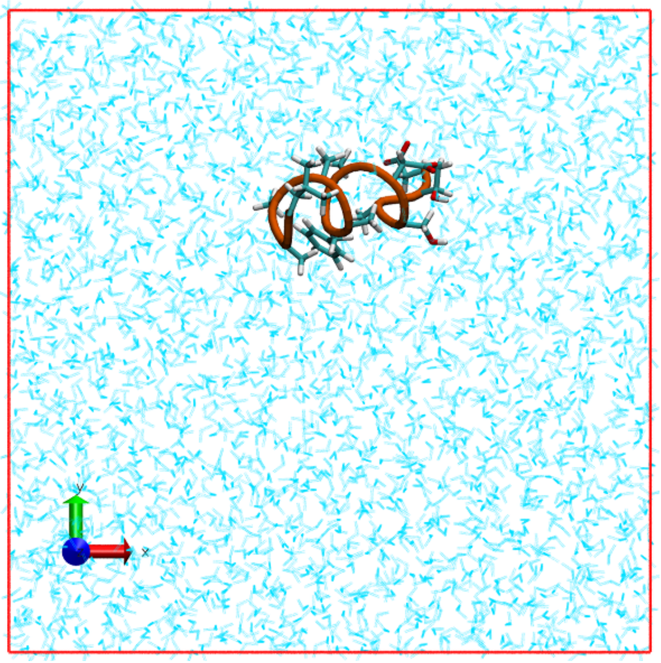


Figure A.2: Final frame of Peptide 1–3 water box simulation.

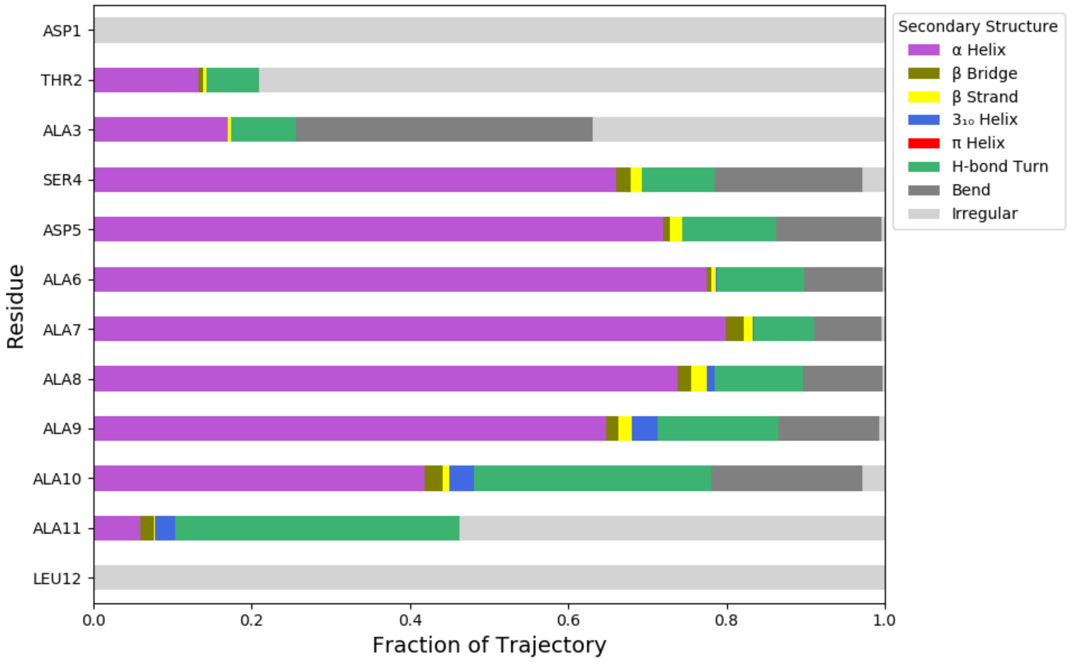


Figure A.3: Secondary structures calculated with DSSP algorithm throughout Peptide 1–1 water box simulation.

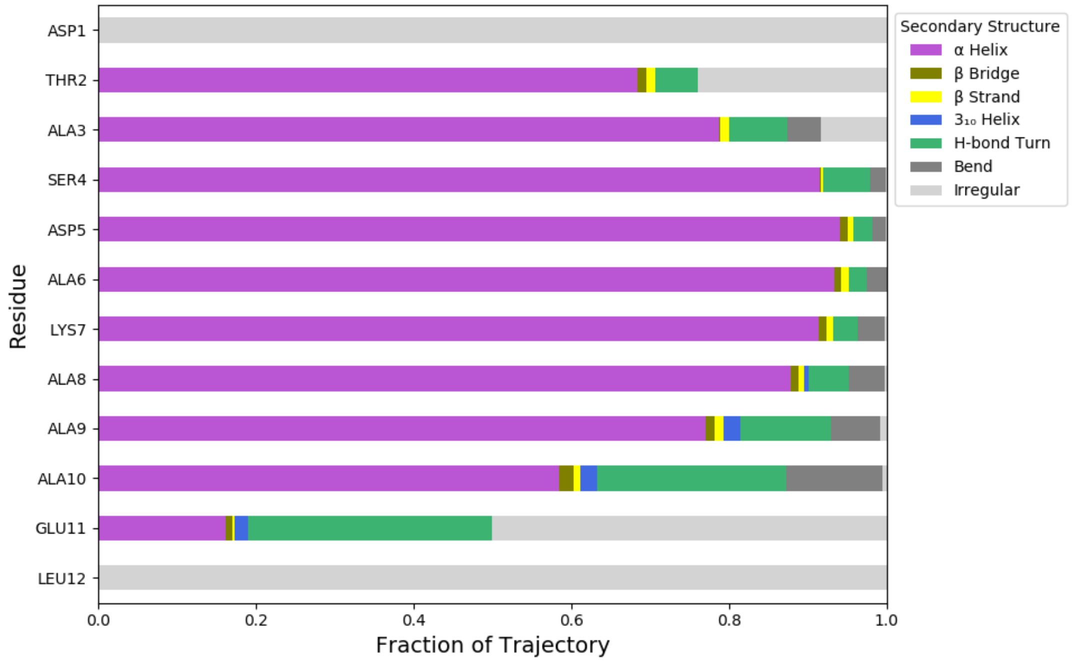


Figure A.4: Secondary structures calculated with DSSP algorithm throughout Peptide 1–2 water box simulation.

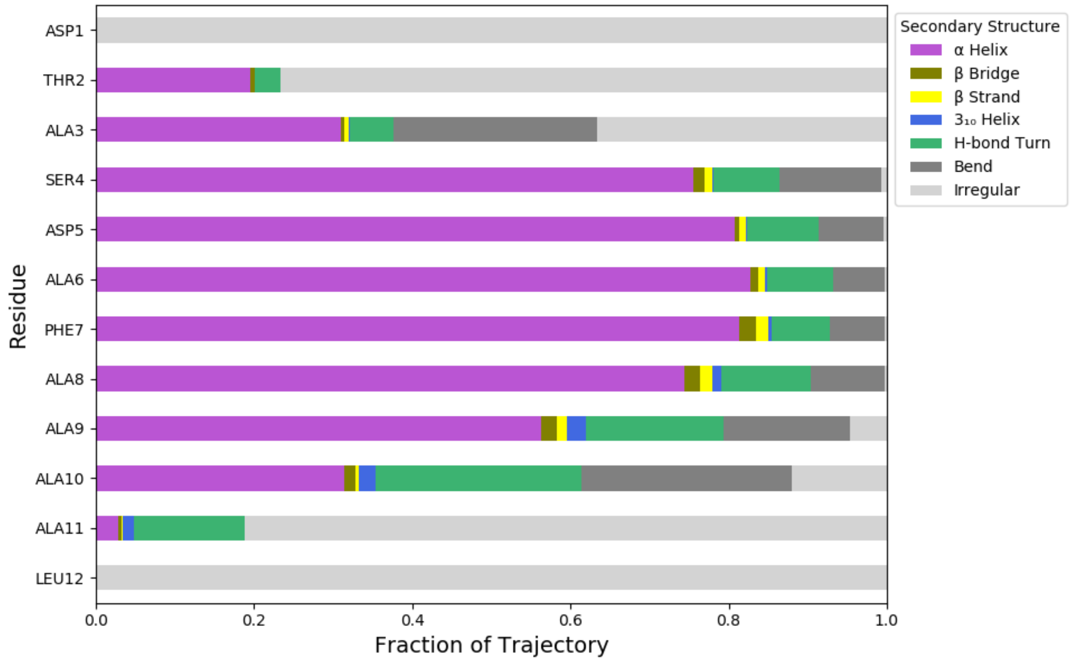


Figure A.5: Secondary structures calculated with DSSP algorithm throughout Peptide 1–3 water box simulation.

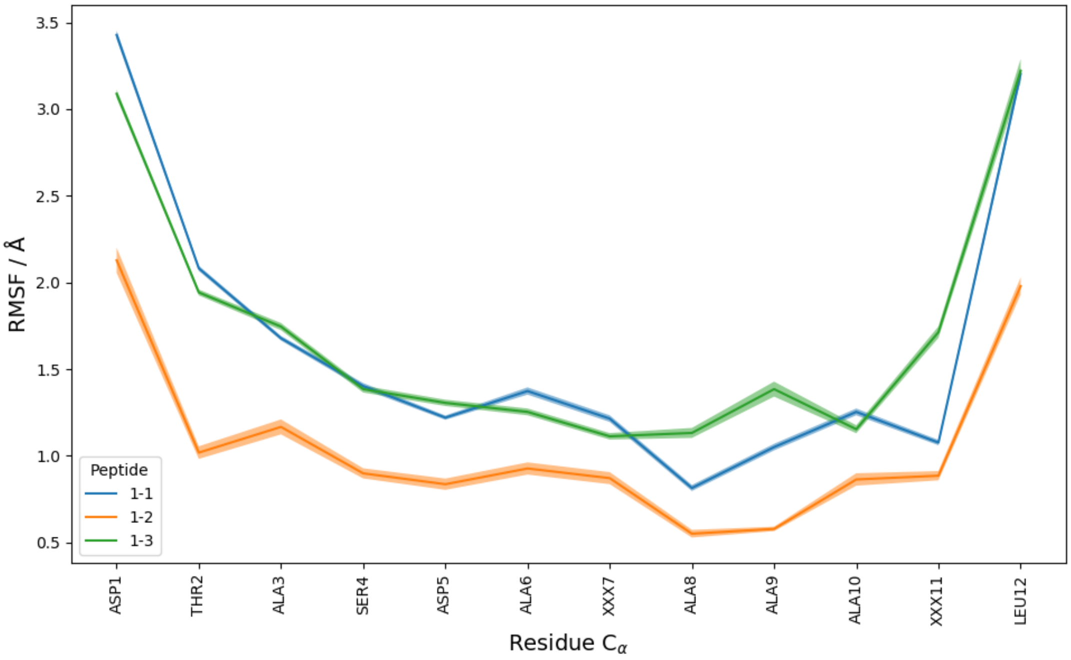


Figure A.6: RMSF calculated on antifreeze peptides water box trajectories.

A.2 Fixed-ice simulations

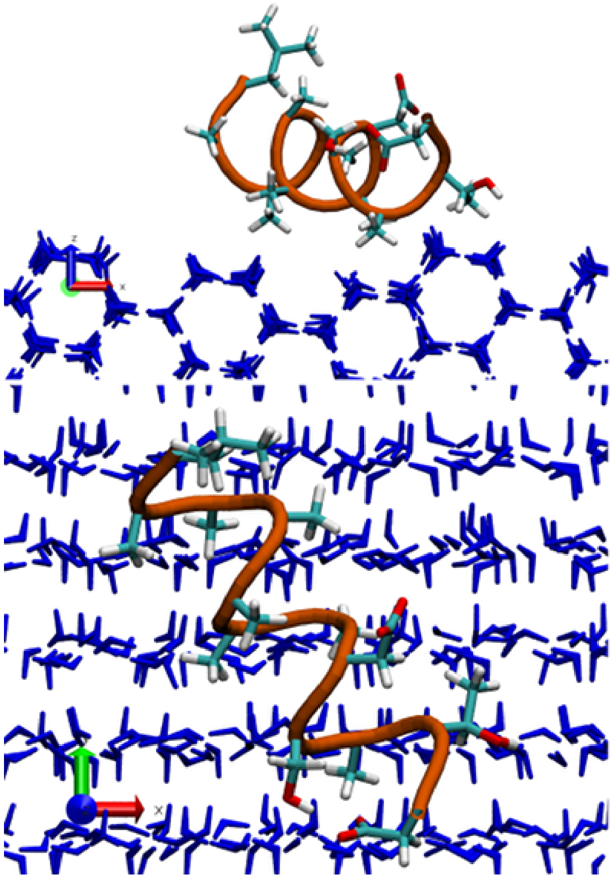


Figure A.7: Views of Peptide 1–1 fixed-ice simulation final frame.

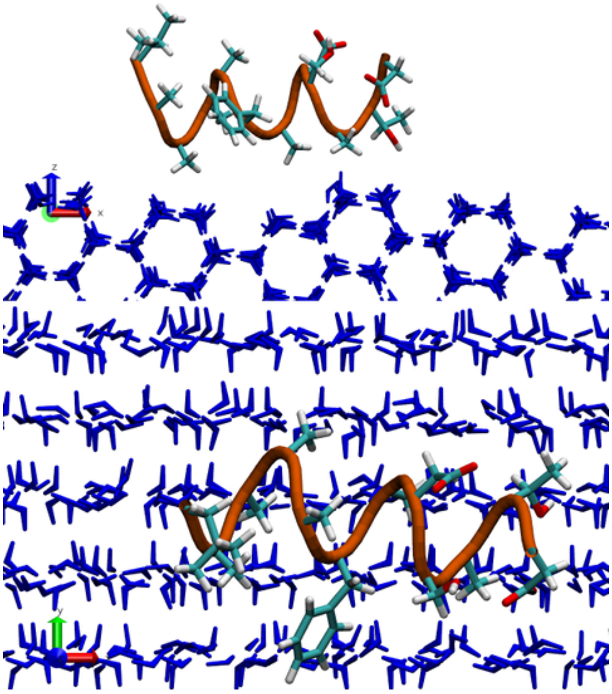


Figure A.8: Views of Peptide 1–3 fixed-ice simulation final frame.

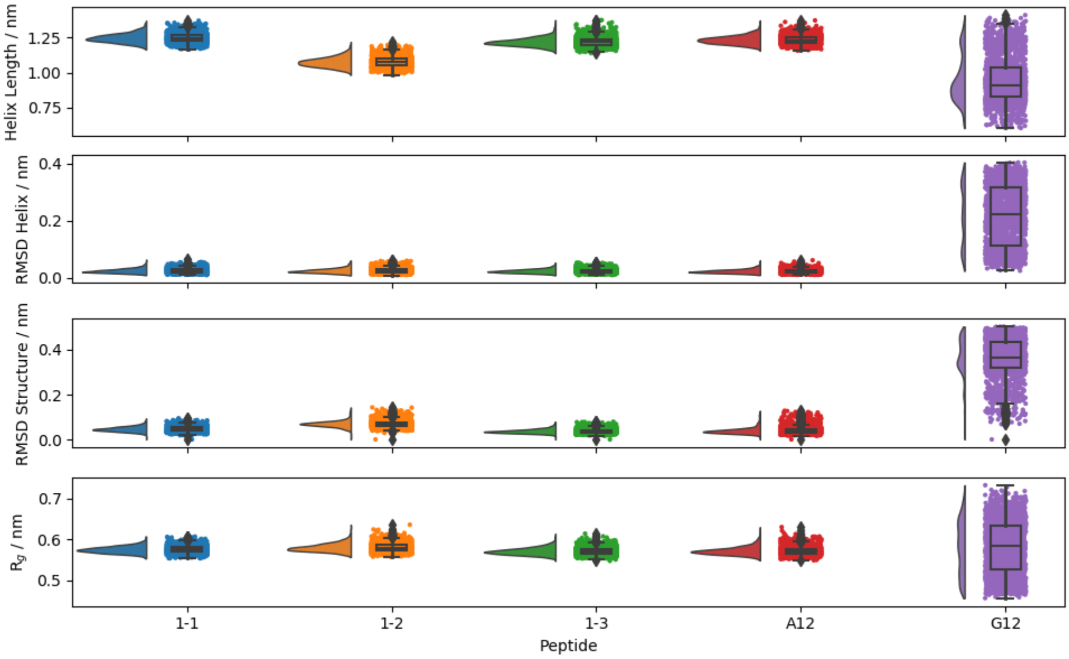


Figure A.9: Distributions of four structural properties calculated throughout the fixed-ice simulations.

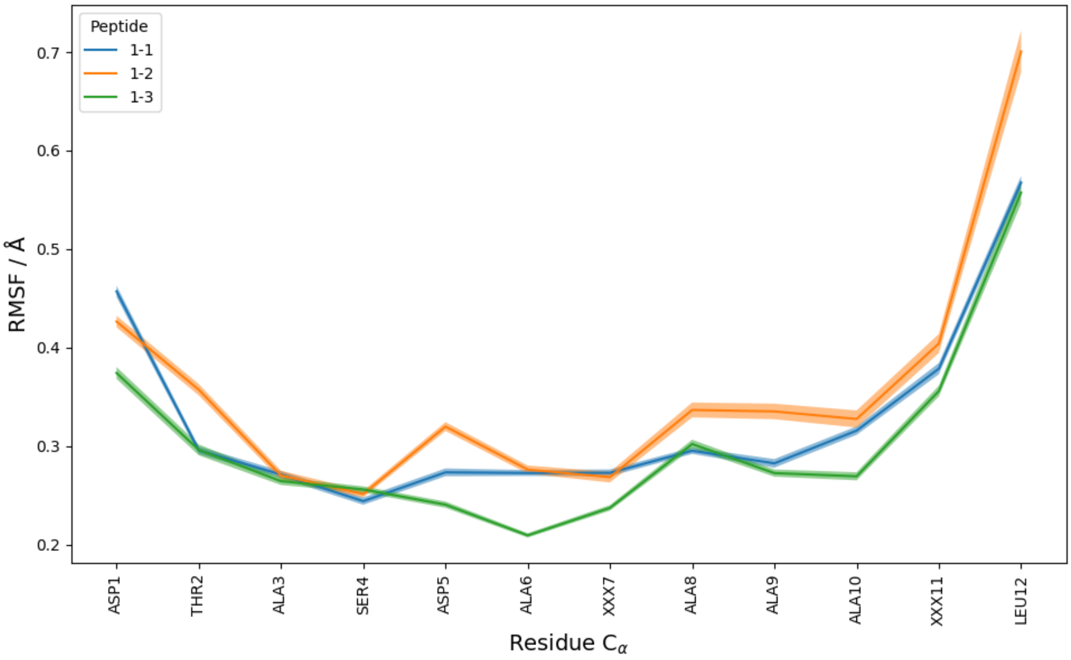


Figure A.10: RMSF calculated on antifreeze peptides fixed-ice trajectories.

A.3 Growing-ice simulations

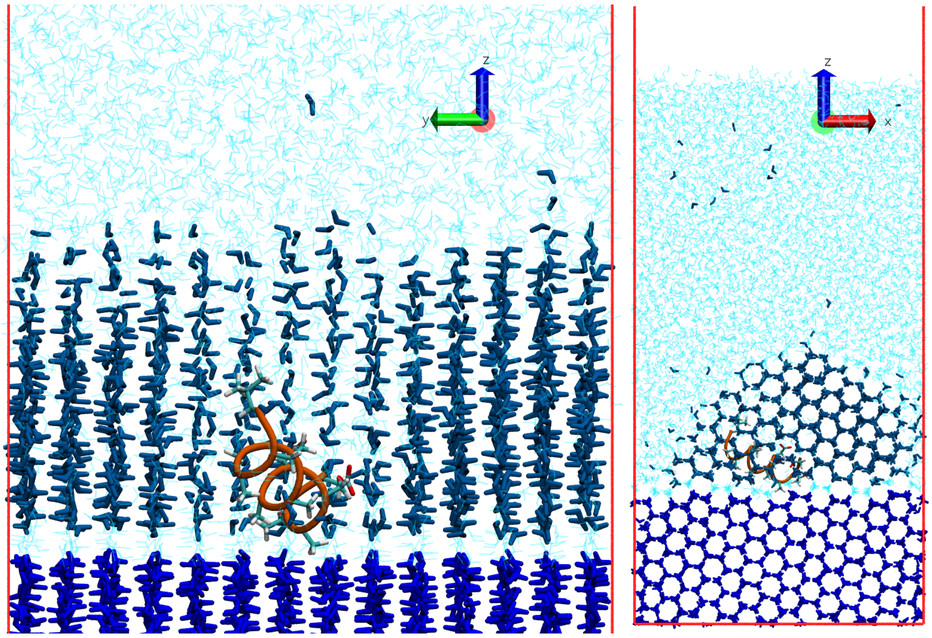


Figure A.11: Views of the final frame of the growing-ice simulation of Peptide 1–1.

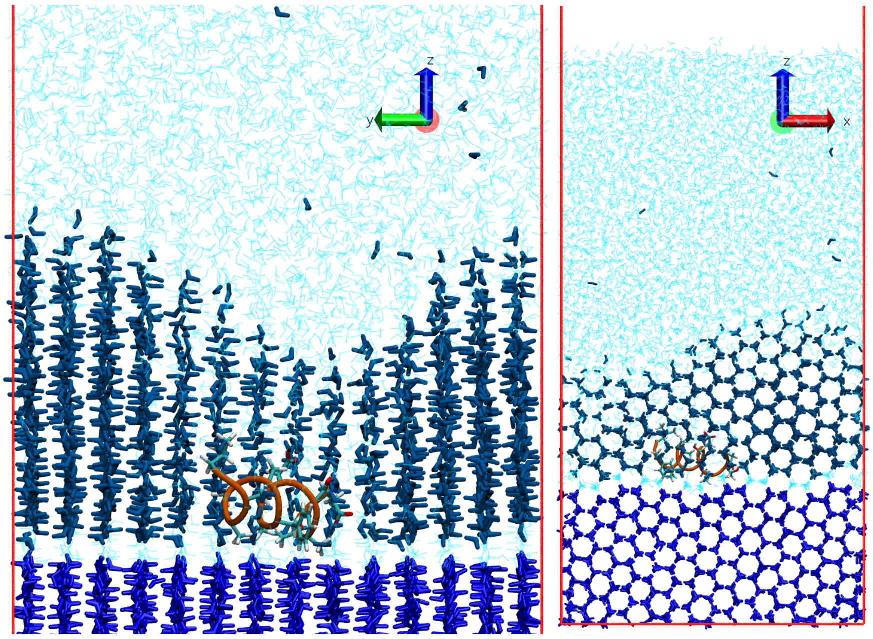


Figure A.12: Views of the final frame of the growing-ice simulation of Peptide 1–3.

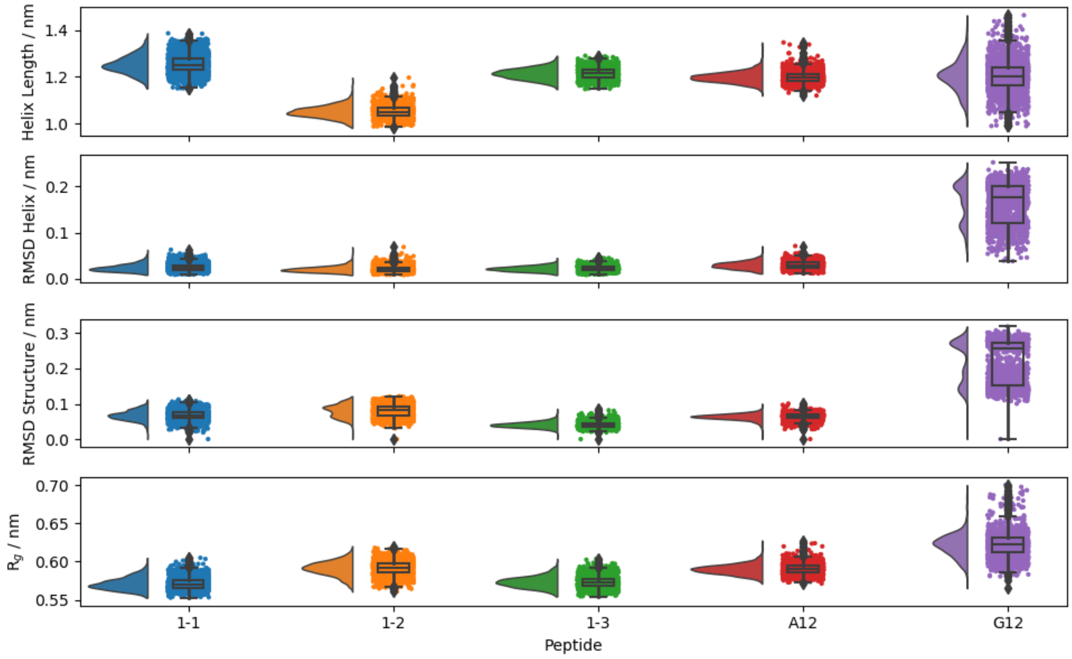


Figure A.13: Distributions of four structural properties calculated throughout the growing-ice simulations.

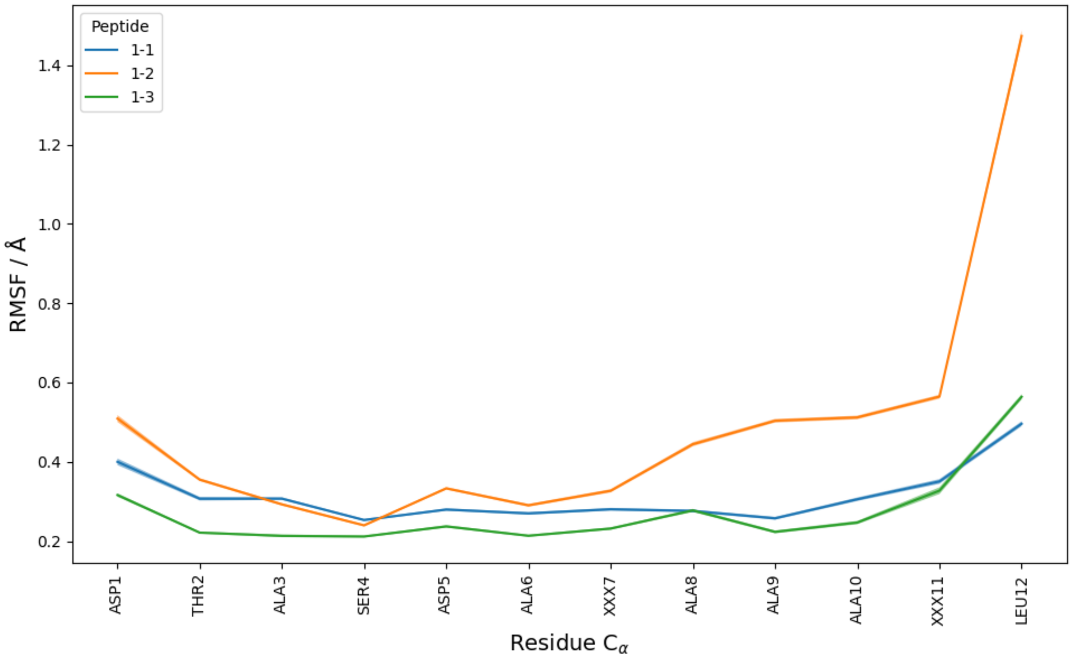


Figure A.14: RMSF calculated on antifreeze peptides growing-ice trajectories.

Appendix B

Detailed survey results

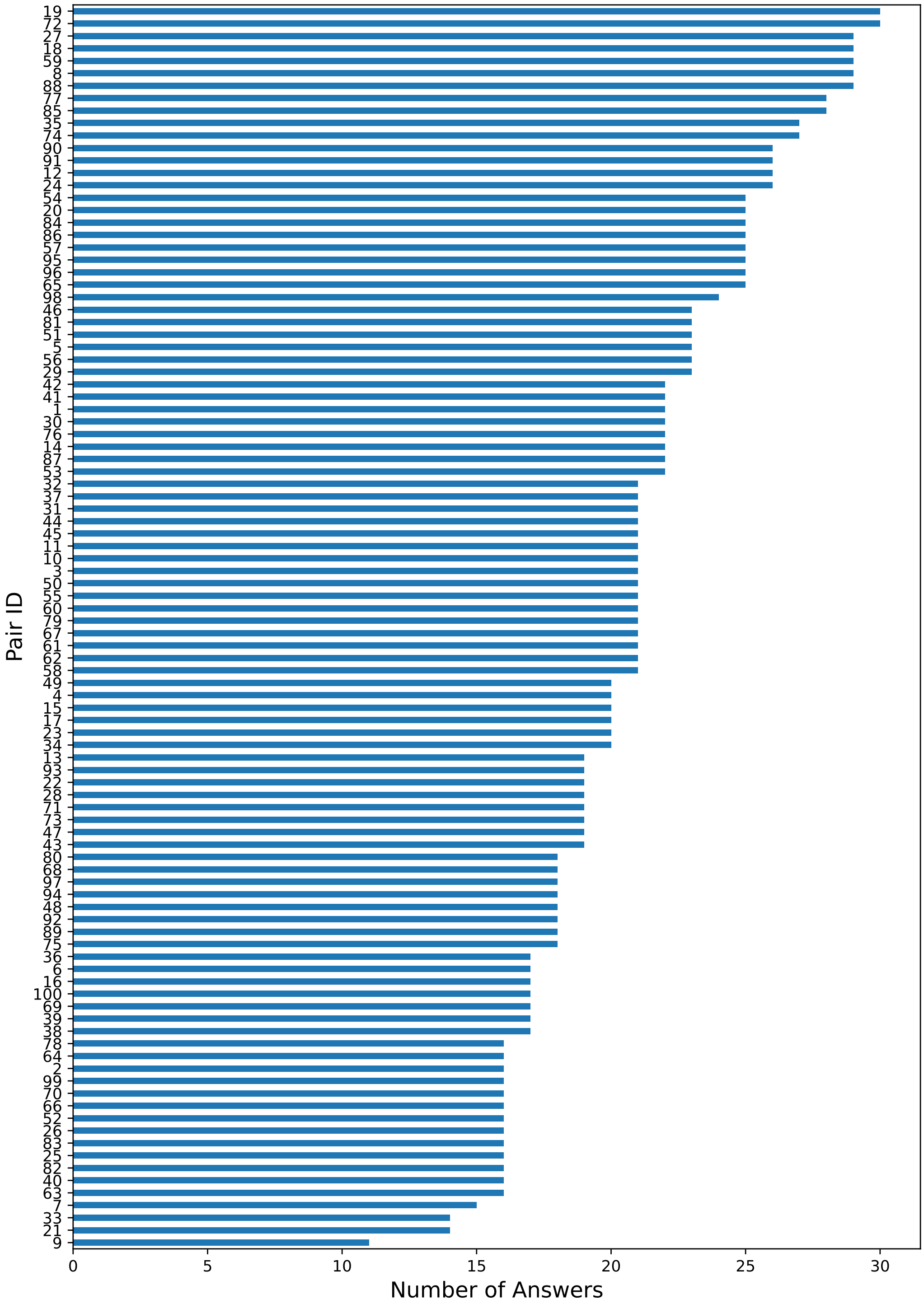


Figure B.1: Number of times each molecule-pair was assessed by survey users.

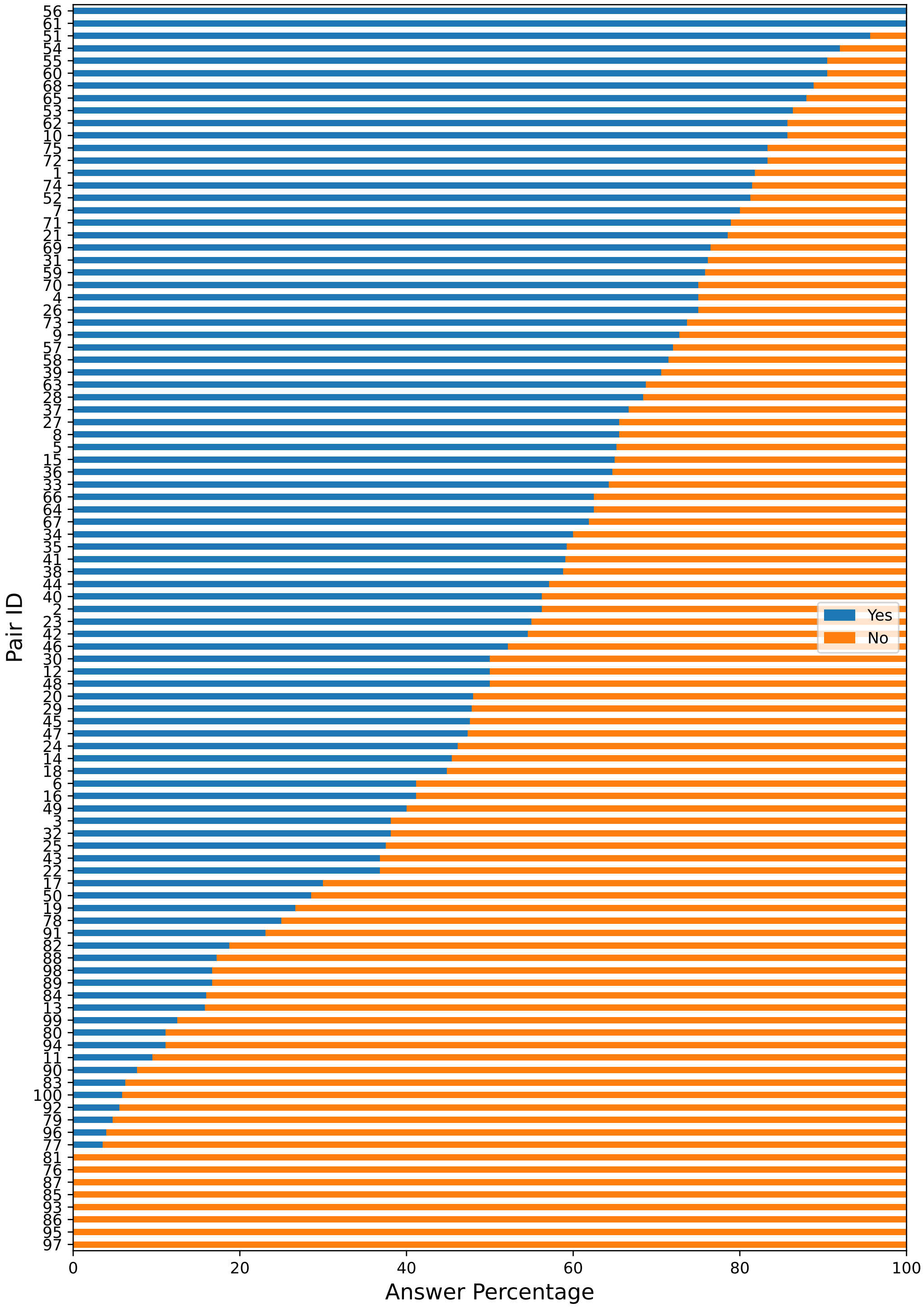


Figure B.2: Percentage of survey users that considered each molecule-pair to be similar or dissimilar.

Appendix C

The new data-set

The new data-set is reported in Table C.1. Each row corresponds to a molecule-pair. We report the Tanimoto CDK Extended and TanimotoCombo similarity measures, the similarity subset (see subsection 8.4.1), the percentage of survey users that considered the molecule-pair to be similar, and the number of answers received by the molecule-pair. The complete new data-set also contains SMILES strings for the two molecules of each molecule-pair. But SMILES strings can be quite long, they would not be formatted nicely, and they are not very useful on printed page or on PDF files. Therefore, SMILES strings are not reported in Table C.1.

The complete new data-set (with SMILES strings and conformers) is available as a CSV file (more useful for cheminformatics) in a GitHub repository:
`https://github.com/enricogandini/paper_similarity_prediction/`
The repository also contains the source code, data, and instructions needed to host the web-app on Heroku and erogate the survey.

Table C.1: The new data-set — calculated similarity measures and assessed similarity percentages.

Pair ID	Tanimoto CDK Extended	Tanimoto Combo	Subset	Similarity Percentage	Number of Answers
1	0.567	1.989	dis2D,sim3D	81.8	22
2	0.532	1.782	dis2D,sim3D	56.2	16
3	0.549	1.778	dis2D,sim3D	38.1	21
4	0.559	1.764	dis2D,sim3D	75.0	20
5	0.453	1.757	dis2D,sim3D	65.2	23
6	0.626	1.757	dis2D,sim3D	41.2	17
7	0.467	1.752	dis2D,sim3D	80.0	15
8	0.522	1.704	dis2D,sim3D	65.5	29

9	0.388	1.674	dis2D,sim3D	72.7	11
10	0.620	1.660	dis2D,sim3D	85.7	21
11	0.275	1.629	dis2D,sim3D	9.5	21
12	0.328	1.627	dis2D,sim3D	50.0	26
13	0.323	1.582	dis2D,sim3D	15.8	19
14	0.482	1.579	dis2D,sim3D	45.5	22
15	0.591	1.562	dis2D,sim3D	65.0	20
16	0.320	1.519	dis2D,sim3D	41.2	17
17	0.422	1.502	dis2D,sim3D	30.0	20
18	0.603	1.460	dis2D,sim3D	44.8	29
19	0.369	1.459	dis2D,sim3D	26.7	30
20	0.608	1.454	dis2D,sim3D	48.0	25
21	0.645	1.454	dis2D,sim3D	78.6	14
22	0.200	1.450	dis2D,sim3D	36.8	19
23	0.349	1.444	dis2D,sim3D	55.0	20
24	0.538	1.437	dis2D,sim3D	46.2	26
25	0.386	1.412	dis2D,sim3D	37.5	16
26	0.878	1.191	sim2D,dis3D	75.0	16
27	0.859	1.292	sim2D,dis3D	65.5	29
28	0.858	1.232	sim2D,dis3D	68.4	19
29	0.848	1.137	sim2D,dis3D	47.8	23
30	0.830	1.271	sim2D,dis3D	50.0	22
31	0.826	1.127	sim2D,dis3D	76.2	21
32	0.802	1.275	sim2D,dis3D	38.1	21
33	0.774	1.060	sim2D,dis3D	64.3	14
34	0.773	0.990	sim2D,dis3D	60.0	20
35	0.772	1.058	sim2D,dis3D	59.3	27
36	0.769	1.092	sim2D,dis3D	64.7	17
37	0.767	1.133	sim2D,dis3D	66.7	21
38	0.765	1.099	sim2D,dis3D	58.8	17
39	0.763	1.052	sim2D,dis3D	70.6	17
40	0.743	1.292	sim2D,dis3D	56.2	16
41	0.740	1.119	sim2D,dis3D	59.1	22
42	0.738	1.090	sim2D,dis3D	54.5	22
43	0.737	1.007	sim2D,dis3D	36.8	19
44	0.717	1.112	sim2D,dis3D	57.1	21
45	0.716	1.142	sim2D,dis3D	47.6	21
46	0.712	1.033	sim2D,dis3D	52.2	23

47	0.707	1.232	sim2D,dis3D	47.4	19
48	0.706	1.011	sim2D,dis3D	50.0	18
49	0.705	1.248	sim2D,dis3D	40.0	20
50	0.701	1.288	sim2D,dis3D	28.6	21
51	1.000	1.980	sim2D,sim3D	95.7	23
52	1.000	1.946	sim2D,sim3D	81.2	16
53	0.966	1.897	sim2D,sim3D	86.4	22
54	0.963	1.826	sim2D,sim3D	92.0	25
55	0.959	1.985	sim2D,sim3D	90.5	21
56	0.959	1.984	sim2D,sim3D	100.0	23
57	0.951	1.897	sim2D,sim3D	72.0	25
58	0.950	1.894	sim2D,sim3D	71.4	21
59	0.947	1.914	sim2D,sim3D	75.9	29
60	0.943	1.976	sim2D,sim3D	90.5	21
61	0.929	1.902	sim2D,sim3D	100.0	21
62	0.923	1.748	sim2D,sim3D	85.7	21
63	0.916	1.915	sim2D,sim3D	68.8	16
64	0.912	1.972	sim2D,sim3D	62.5	16
65	0.909	1.812	sim2D,sim3D	88.0	25
66	0.904	1.565	sim2D,sim3D	62.5	16
67	0.892	1.700	sim2D,sim3D	61.9	21
68	0.888	1.864	sim2D,sim3D	88.9	18
69	0.878	1.838	sim2D,sim3D	76.5	17
70	0.875	1.657	sim2D,sim3D	75.0	16
71	0.872	1.733	sim2D,sim3D	78.9	19
72	0.870	1.716	sim2D,sim3D	83.3	30
73	0.860	1.752	sim2D,sim3D	73.7	19
74	0.811	1.742	sim2D,sim3D	81.5	27
75	0.760	1.738	sim2D,sim3D	83.3	18
76	0.179	0.443	dis2D,dis3D	0.0	22
77	0.184	0.564	dis2D,dis3D	3.6	28
78	0.193	0.696	dis2D,dis3D	25.0	16
79	0.206	0.589	dis2D,dis3D	4.8	21
80	0.216	0.837	dis2D,dis3D	11.1	18
81	0.230	0.894	dis2D,dis3D	0.0	23
82	0.240	0.760	dis2D,dis3D	18.8	16
83	0.241	0.862	dis2D,dis3D	6.2	16
84	0.242	0.934	dis2D,dis3D	16.0	25

85	0.250	0.529	dis2D,dis3D	0.0	28
86	0.254	0.777	dis2D,dis3D	0.0	25
87	0.268	0.780	dis2D,dis3D	0.0	22
88	0.288	0.674	dis2D,dis3D	17.2	29
89	0.290	0.886	dis2D,dis3D	16.7	18
90	0.292	0.870	dis2D,dis3D	7.7	26
91	0.299	0.888	dis2D,dis3D	23.1	26
92	0.302	0.769	dis2D,dis3D	5.6	18
93	0.305	0.796	dis2D,dis3D	0.0	19
94	0.333	0.758	dis2D,dis3D	11.1	18
95	0.335	0.750	dis2D,dis3D	0.0	25
96	0.400	0.741	dis2D,dis3D	4.0	25
97	0.401	0.625	dis2D,dis3D	0.0	18
98	0.406	0.657	dis2D,dis3D	16.7	24
99	0.414	0.762	dis2D,dis3D	12.5	16
100	0.546	0.922	dis2D,dis3D	5.9	17

References

[1] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren, "Molecular dynamics simulations," *Current Opinion in Structural Biology*, vol. 12, no. 2, pp. 190–196, Apr. 2002, ISSN: 0959-440X. DOI: 10.1016/S0959-440X(02)00308-1.

[2] K. Binder, J. Horbach, W. Kob, W. Paul, and F. Varnik, "Molecular dynamics simulations," *Journal of Physics: Condensed Matter*, vol. 16, no. 5, S429–S453, Jan. 2004, ISSN: 0953-8984. DOI: 10.1088/0953-8984/16/5/006.

[3] M. A. González, "Force fields and molecular dynamics simulations," *École thématique de la Société Française de la Neutronique*, vol. 12, pp. 169–200, 2011, ISSN: 2107-7223, 2107-7231. DOI: 10.1051/sfn/201112009.

[4] E. Paquet and H. L. Viktor, "Molecular Dynamics, Monte Carlo Simulations, and Langevin Dynamics: A Computational Review," *BioMed Research International*, vol. 2015, e183918, Feb. 2015, ISSN: 2314-6133. DOI: 10.1155/2015/183918.

[5] A. R. Leach, *Molecular Modelling: Principles and Applications*, Second. Pearson education, 2001.

[6] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed, ser. Computational Science Series 1. San Diego: Academic Press, 2002, ISBN: 978-0-12-267351-1.

[7] B. Leimkuhler and C. Matthews, *Molecular Dynamics*, ser. Interdisciplinary Applied Mathematics. Cham: Springer International Publishing, 2015, vol. 39, ISBN: 978-3-319-16374-1 978-3-319-16375-8. DOI: 10.1007/978-3-319-16375-8.

[8] E. Fermi, J. R. Pasta, and S. Ulam, "Studies of nonlinear problems," Los Alamos Scientific Laboratory, Los Alamos, New Mexico, Technical Report LA-1940, May 1955.

[9] J. De Tullio, "The Fermi-Pasta-Ulam model: The birth of numerical simulation," *Lettera Matematica*, vol. 4, no. 1, pp. 41–48, Mar. 2016, ISSN: 2281-5937. DOI: 10.1007/s40329-016-0126-4.

[10] L. Monticelli and E. Salonen, Eds., *Biomolecular Simulations*, ser. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2013, vol. 924, ISBN: 978-1-62703-016-8 978-1-62703-017-5. DOI: 10.1007/978-1-62703-017-5.

[11] M. Bonomi and C. Camilloni, *Biomolecular Simulations: Methods and Protocols*. 2019, ISBN: 978-1-4939-9608-7. [Online]. Available: https://doi.org/10.1007/978-1-4939-9608-7 (visited on 12/16/2019).

[12] A. Kukol and J. M. Walker, Eds., *Molecular Modeling of Proteins*, ser. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2008, vol. 443, ISBN: 978-1-58829-864-5 978-1-59745-177-2. DOI: 10.1007/978-1-59745-177-2.

[13] A. Kukol, Ed., *Molecular Modeling of Proteins*, Second, ser. Methods in Molecular Biology. New York, NY: Springer New York, 2015, vol. 1215, ISBN: 978-1-4939-1464-7 978-1-4939-1465-4. DOI: 10.1007/978-1-4939-1465-4.

[14] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "GROMACS: Fast, flexible, and free," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[15] O. Guvench and A. D. MacKerell Jr., "Chapter 4 - Comparison of Protein Force Fields for Molecular Dynamics Simulations," in *Molecular Modeling of Proteins*, ser. Methods Molecular Biology™, A. Kukol, Ed., Totowa, NJ: Humana Press, 2008, pp. 63–88, ISBN: 978-1-59745-177-2. DOI: 10.1007/978-1-59745-177-2_4.

[16] L. Monticelli and D. P. Tieleman, "Chapter 8 - Force Fields for Classical Molecular Dynamics," in *Biomolecular Simulations: Methods and Protocols*, ser. Methods in Molecular Biology, L. Monticelli and E. Salonen, Eds., Totowa, NJ: Humana Press, 2013, pp. 197–213, ISBN: 978-1-62703-017-5. DOI: 10.1007/978-1-62703-017-5_8.

[17] P. E. M. Lopes, O. Guvench, and A. D. MacKerell Jr., "Chapter 3 - Current Status of Protein Force Fields for Molecular Dynamics Simulations," in *Molecular Modeling of Proteins*, ser. Methods in Molecular Biology, A. Kukol, Ed., New York, NY: Springer, 2015, pp. 47–71, ISBN: 978-1-4939-1465-4. DOI: 10.1007/978-1-4939-1465-4_3.

[18] N. Foloppe and A. D. MacKerell Jr., "All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data," *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 86–104, 2000, ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(20000130)21:2⟨86::AID-JCC2⟩3.0.CO;2-G.

[19] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, Jul. 1983, ISSN: 0021-9606. DOI: 10.1063/1.445869.

[20] D. J. Price and C. L. Brooks, "A modified TIP3P water potential for simulation with Ewald summation," *The Journal of Chemical Physics*, vol. 121, no. 20, pp. 10 096–10 103, Nov. 2004, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1808117.

[21] C. Vega, J. L. F. Abascal, M. M. Conde, and J. L. Aragones, "What ice can teach us about water interactions: A critical comparison of the performance of different water models," *Faraday Discussions*, vol. 141, no. 0, pp. 251–276, 2009. DOI: 10.1039/B805531A.

[22] J. L. F. Abascal, E. Sanz, R. García Fernández, and C. Vega, "A potential model for the study of ices and amorphous water: TIP4P/Ice," *The Journal of Chemical Physics*, vol. 122, no. 23, p. 234 511, Jun. 2005, ISSN: 0021-9606. DOI: 10.1063/1.1931662.

[23] S. Venketesh and C. Dayananda, "Properties, Potentials, and Prospects of Antifreeze Proteins," *Critical Reviews in Biotechnology*, vol. 28, no. 1, pp. 57–82, Jan. 2008, ISSN: 0738-8551. DOI: 10.1080/07388550801891152.

[24] M. Bar Dolev, I. Braslavsky, and P. L. Davies, "Ice-Binding Proteins and Their Function," *Annual Review of Biochemistry*, vol. 85, no. 1, pp. 515–542, Jun. 2016, ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev-biochem-060815-014546.

[25] P. L. Davies, "Ice-binding proteins: A remarkable diversity of structures for stopping and starting ice growth," *Trends in Biochemical Sciences*, vol. 39, no. 11, pp. 548–555, Nov. 2014, ISSN: 0968-0004. DOI: 10.1016/j.tibs.2014.09.005.

[26] M. M. Harding, L. G. Ward, and A. D. J. Haymet, "Type I 'antifreeze' proteins," *European Journal of Biochemistry*, vol. 264, no. 3, pp. 653–665, 1999, ISSN: 1432-1033. DOI: 10.1046/j.1432-1327.1999.00617.x.

[27] F. Sicheri and D. S. C. Yang, "Ice-binding structure and mechanism of an antifreeze protein from winter flounder," *Nature*, vol. 375, pp. 427–431, 1995.

[28] A. Wierzbicki, P. Dalal, T. E. Cheatham, J. E. Knickelbein, A. D. J. Haymet, and J. D. Madura, "Antifreeze Proteins at the Ice/Water Interface: Three Calculated Discriminating Properties for Orientation of Type I Proteins," *Biophysical Journal*, vol. 93, no. 5, pp. 1442–1451, Sep. 2007, ISSN: 0006-3495. DOI: 10.1529/biophysj.107.105189.

[29] E. Gandini, M. Sironi, and S. Pieraccini, “Modelling of short synthetic antifreeze peptides: Insights into ice-pinning mechanism,” *Journal of Molecular Graphics and Modelling*, vol. 100, p. 107 680, Nov. 2020, ISSN: 1093-3263. DOI: 10.1016/j.jmgm.2020.107680.

[30] P. L. Davies, Jason Baardsnes, M. J. Kuiper, and V. K. Walker, “Structure and function of antifreeze proteins,” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 357, no. 1423, D. J. Bowles, P. J. Lillford, D. A. Rees, and I. A. Shanks, Eds., pp. 927–935, Jul. 2002. DOI: 10.1098/rstb.2002.1081.

[31] L. L. C. Olijve, K. Meister, A. L. DeVries, J. G. Duman, S. Guo, H. J. Bakker, and I. K. Voets, “Blocking rapid ice crystal growth through nonbasal plane adsorption of antifreeze proteins,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 14, pp. 3740–3745, Apr. 2016, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1524109113.

[32] C. A. Knight, D. Wen, and R. A. Laursen, “Nonequilibrium Antifreeze Peptides and the Recrystallization of Ice,” *Cryobiology*, vol. 32, no. 1, pp. 23–34, Feb. 1995, ISSN: 0011-2240. DOI: 10.1006/cryo.1995.1002.

[33] P. Wilson, “Explaining thermal hysteresis by the Kelvin effect,” *CryoLetters*, vol. 14, pp. 31–36, 1993.

[34] I. K. Voets, “From ice-binding proteins to bio-inspired antifreeze materials,” *Soft Matter*, vol. 13, no. 28, pp. 4808–4823, 2017. DOI: 10.1039/C6SM02867E.

[35] Y. Zhang, K. Liu, K. Li, V. Gutowski, Y. Yin, and J. Wang, “Fabrication of Anti-Icing Surfaces by Short α-Helical Peptides,” *ACS Applied Materials & Interfaces*, vol. 10, no. 2, pp. 1957–1962, Jan. 2018, ISSN: 1944-8244, 1944-8252. DOI: 10.1021/acsami.7b13130.

[36] H. Nada and Y. Furukawa, “Antifreeze proteins: Computer simulation studies on the mechanism of ice growth inhibition,” *Polymer Journal*, vol. 44, no. 7, pp. 690–698, Jul. 2012, ISSN: 1349-0540. DOI: 10.1038/pj.2012.13.

[37] R. K. Kar and A. Bhunia, “Biophysical and biochemical aspects of antifreeze proteins: Using computational tools to extract atomistic information,” *Progress in Biophysics and Molecular Biology*, vol. 119, no. 2, pp. 194–204, Nov. 2015, ISSN: 00796107. DOI: 10.1016/j.pbiomolbio.2015.09.001.

[38] A. Hudait, D. R. Moberg, Y. Qiu, N. Odendahl, F. Paesani, and V. Molinero, “Preordering of water is not needed for ice recognition by hyperactive antifreeze proteins,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, pp. 8266–8271, Aug. 2018, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1806996115.

[39] A. Hudait, Y. Qiu, N. Oden, and V. Molinero, "Hydrogen-Bonding and Hydrophobic Groups Contribute Equally to the Binding of Hyperactive Antifreeze and Ice-Nucleating Proteins to Ice — Journal of the American Chemical Society," *Journal of American Chemical Society*, vol. 141, no. 19, pp. 7887–7898, 2019. DOI: 10.1021/jacs.9b02248.

[40] M. J. Kuiper, C. J. Morton, S. E. Abraham, and A. Gray-Weale, "The biological function of an insect antifreeze protein simulated by molecular dynamics," *eLife*, vol. 4, May 2015, ISSN: 2050-084X. DOI: 10.7554/eLife.05142.

[41] S. S. Mallajosyula, K. Vanommeslaeghe, and A. D. MacKerell Jr., "Perturbation of Long-Range Water Dynamics as the Mechanism for the Antifreeze Activity of Antifreeze Glycoprotein," *The Journal of Physical Chemistry B*, vol. 118, no. 40, pp. 11 696–11 706, Oct. 2014, ISSN: 1520-6106, 1520-5207. DOI: 10.1021/jp508128d.

[42] K. Meister, S. Strazdaite, A. L. DeVries, S. Lotze, L. L. C. Olijve, I. K. Voets, and H. J. Bakker, "Observation of ice-like water layers at an aqueous protein surface," *Proceedings of the National Academy of Sciences*, vol. 111, no. 50, pp. 17 732–17 736, Dec. 2014, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1414188111.

[43] S. Ebbinghaus, K. Meister, B. Born, A. L. DeVries, M. Gruebele, and M. Havenith, "Antifreeze Glycoprotein Activity Correlates with Long-Range Protein-Water Dynamics," *Journal of the American Chemical Society*, vol. 132, no. 35, pp. 12 210–12 211, Sep. 2010, ISSN: 0002-7863. DOI: 10.1021/ja1051632.

[44] R. E. Feeney and Y. Yeh, "Antifreeze proteins: Current status and possible food uses," *Trends in Food Science & Technology*, vol. 9, no. 3, pp. 102–106, Mar. 1998, ISSN: 0924-2244. DOI: 10.1016/S0924-2244(98)00025-9.

[45] N. S. Ustun and S. Turhan, "Antifreeze Proteins: Characteristics, Function, Mechanism of Action, Sources and Application to Foods," *Journal of Food Processing and Preservation*, vol. 39, no. 6, pp. 3189–3197, 2015, ISSN: 1745-4549. DOI: 10.1111/jfpp.12476.

[46] P. Dalal and F. D. Sönnichsen, "Source of the Ice-Binding Specificity of Antifreeze Protein Type I," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 5, pp. 1276–1284, Sep. 2000, ISSN: 0095-2338. DOI: 10.1021/ci000449b.

[47] S. Chakraborty and B. Jana, "Conformational and hydration properties modulate ice recognition by type I antifreeze protein and its mutants," *Physical Chemistry Chemical Physics*, vol. 19, no. 18, pp. 11 678–11 689, May 2017, ISSN: 1463-9084. DOI: 10.1039/C7CP00221A.

[48] T. Nobekawa and Y. Hagiwara, "Interaction among the twelve-residue segment of antifreeze protein type I, or its mutants, water and a hexagonal ice crystal," *Molecular Simulation*, vol. 34, no. 6, pp. 591–610, May 2008, ISSN: 0892-7022. DOI: 10.1080/08927020801986556.

[49] R. K. Kar and A. Bhunia, "Will It Be Beneficial To Simulate the Antifreeze Proteins at Ice Freezing Condition or at Lower Temperature?" *The Journal of Physical Chemistry B*, vol. 119, no. 35, pp. 11 485–11 495, Sep. 2015, ISSN: 1520-6106, 1520-5207. DOI: 10.1021/acs.jpcb.5b04919.

[50] D. Wen and R. A. Laursen, "A model for binding of an antifreeze polypeptide to ice," *Biophysical Journal*, vol. 63, no. 6, pp. 1659–1662, Dec. 1992, ISSN: 0006-3495. DOI: 10.1016/S0006-3495(92)81750-2.

[51] C. A. Knight, C. C. Cheng, and A. L. DeVries, "Adsorption of alpha-helical antifreeze peptides on specific ice crystal surface planes," *Biophysical Journal*, vol. 59, no. 2, pp. 409–418, Feb. 1991, ISSN: 0006-3495. DOI: 10.1016/S0006-3495(91)82234-2.

[52] F. Sicheri and D. S. C. Yang, "Structure Determination of a Lone α-Helical Antifreeze Protein from Winter Flounder," *Acta Crystallographica Section D: Biological Crystallography*, vol. 52, no. 3, pp. 486–498, May 1996, ISSN: 0907-4449. DOI: 10.1107/S0907444995015253.

[53] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera - A visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004, ISSN: 1096-987X. DOI: 10.1002/jcc.20084.

[54] M. V. Shapovalov and R. L. Dunbrack, "A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions," *Structure*, vol. 19, no. 6, pp. 844–858, Jun. 2011, ISSN: 0969-2126. DOI: 10.1016/j.str.2011.03.019.

[55] H. Lee, "Structures, dynamics, and hydrogen-bond interactions of antifreeze proteins in TIP4P/Ice water and their dependence on force fields," *PLOS ONE*, vol. 13, no. 6, e0198887, Jun. 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0198887.

[56] M. Matsumoto, T. Yagasaki, and H. Tanaka, "GenIce: Hydrogen-Disordered Ice Generator," *Journal of Computational Chemistry*, vol. 39, no. 1, pp. 61–64, 2018, ISSN: 1096-987X. DOI: 10.1002/jcc.25077.

[57] K. Momma and F. Izumi, "VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data," *Journal of Applied Crystallography*, vol. 44, no. 6, pp. 1272–1276, Dec. 2011, ISSN: 0021-8898. DOI: 10.1107/S0021889811038970.

[58] K. Mochizuki and V. Molinero, "Antifreeze Glycoproteins Bind Reversibly to Ice via Hydrophobic Groups," *Journal of the American Chemical Society*, vol. 140, no. 14, pp. 4803–4811, Apr. 2018, ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.7b13630.

[59] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, Feb. 1996, ISSN: 0263-7855. DOI: 10.1016/0263-7855(96)00018-5.

[60] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983, ISSN: 1097-0282. DOI: 10.1002/bip.360221211.

[61] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories," *Biophysical Journal*, vol. 109, no. 8, pp. 1528–1532, Oct. 2015, ISSN: 0006-3495. DOI: 10.1016/j.bpj.2015.08.015.

[62] R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. ski, D. L. Dotson, S. Buchoux, I. M. Kenney, and O. Beckstein, "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations," in *Proceedings of the 15th Python in Science Conference*, 2016, p. 8.

[63] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations," *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011, ISSN: 1096-987X. DOI: 10.1002/jcc.21787.

[64] A. H. Nguyen and V. Molinero, "Identification of Clathrate Hydrates, Hexagonal Ice, Cubic Ice, and Liquid Water in Simulations: The CHILL+ Algorithm," *The Journal of Physical Chemistry B*, vol. 119, no. 29, pp. 9369–9376, Jul. 2015, ISSN: 1520-6106. DOI: 10.1021/jp510289t.

[65] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," *Computing in Science Engineering*, vol. 9, no. 3, pp. 90–95, May 2007, ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[66] *Seaborn: Statistical Data Visualization*, Feb. 2020. [Online]. Available: https://seaborn.pydata.org/ (visited on 02/12/2020).

[67] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, and R. A. Kievit, "Raincloud plots: A multi-platform tool for robust data visualization," *Wellcome Open Research*, vol. 4, p. 63, Apr. 2019, ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.15191.2.

[68] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[69] H. Abdi and L. J. Williams, "Principal component Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, ISSN: 19395108. DOI: 10.1002/wics.101.

[70] M. Tiberti, E. Papaleo, T. Bengtsen, W. Boomsma, and K. Lindorff-Larsen, "ENCORE: Software for Quantitative Ensemble Comparison," *PLOS Computational Biology*, vol. 11, no. 10, B. L. de Groot, Ed., e1004415, Oct. 2015, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004415.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 06/12/2020).

[72] A. Mitra and D. Sept, "Taxol Allosterically Alters the Dynamics of the Tubulin Dimer and Increases the Flexibility of Microtubules," *Biophysical Journal*, vol. 95, no. 7, pp. 3252–3258, Oct. 2008, ISSN: 0006-3495. DOI: 10.1529/biophysj.108.133884.

[73] F. Dapiaggi, S. Pieraccini, and M. Sironi, "In silico study of VP35 inhibitors: From computational alanine scanning to essential dynamics," *Molecular BioSystems*, vol. 11, no. 8, pp. 2152–2157, 2015, ISSN: 1742-206X, 1742-2051. DOI: 10.1039/C5MB00348B.

[74] H. Kun and Y. Mastai, "Activity of short segments of Type I antifreeze protein," *Peptide Science*, vol. 88, no. 6, pp. 807–814, 2007, ISSN: 1097-0282. DOI: 10.1002/bip.20844.

[75] P. M. Naullage, Y. Qiu, and V. Molinero, "What Controls the Limit of Supercooling and Superheating of Pinned Ice Surfaces?" *The Journal of Physical Chemistry Letters*, vol. 9, no. 7, pp. 1712–1720, Apr. 2018, ISSN: 1948-7185. DOI: 10.1021/acs.jpclett.8b00300.

[76] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal Component Analysis for Protein Folding Dynamics," *Journal of Molecular Biology*, vol. 385, no. 1, pp. 312–329, Jan. 2009, ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.10.018.

[77] H. W. Ng, C. A. Laughton, and S. W. Doughty, "Molecular Dynamics Simulations of the Adenosine A2a Receptor: Structural Stability, Sampling, and Convergence," *Journal of Chemical Information and Modeling*, vol. 53, no. 5, pp. 1168–1178, May 2013, ISSN: 1549-9596. DOI: 10.1021/ci300610w.

[78] C. C. David and D. J. Jacobs, "Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins," *Methods in molecular biology (Clifton, N.J.)*, vol. 1084, pp. 193–226, 2014, ISSN: 1064-3745. DOI: 10.1007/978-1-62703-658-0_11.

[79] S. Ebbinghaus, K. Meister, M. B. Prigozhin, A. L. DeVries, M. Havenith, J. Dzubiella, and M. Gruebele, "Functional Importance of Short-Range Binding and Long-Range Solvent Interactions in Helical Antifreeze Peptides," *Biophysical Journal*, vol. 103, no. 2, pp. L20–L22, Jul. 2012, ISSN: 0006-3495. DOI: 10.1016/j.bpj.2012.06.013.

[80] K. Meister, S. Ebbinghaus, Y. Xu, J. G. Duman, A. DeVries, M. Gruebele, D. M. Leitner, and M. Havenith, "Long-range protein–water dynamics in hyperactive insect antifreeze proteins," *Proceedings of the National Academy of Sciences*, vol. 110, no. 5, pp. 1617–1622, Jan. 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1214911110.

[81] J. Lee, S. Y. Lee, D.-K. Lim, D. J. Ahn, and S. Lee, "Antifreezing Gold Colloids," *Journal of the American Chemical Society*, vol. 141, no. 47, pp. 18 682–18 693, Nov. 2019, ISSN: 0002-7863, 1520-5126. DOI: 10.1021/jacs.9b05526.

[82] C. G. Paris, *Extract from the 218th ACS National Meeting and Exposition, quoted by W. Warr at https://www.warr.com/warrzone2000.html*, Aug. 1999.

[83] M. Hann and R. Green, "Chemoinformatics - a new name for an old problem?" *Current Opinion in Chemical Biology*, vol. 3, no. 4, pp. 379–383, Aug. 1999, ISSN: 1367-5931. DOI: 10.1016/S1367-5931(99)80057-X.

[84] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*, Revised Edition. Dordrecht: Springer, 2007, ISBN: 978-1-4020-6290-2.

[85] B. A. Bunin, B. Siesel, G. Morales, and J. Bajorath, *Chemoinformatics: Theory, Practice, and Products*. Springer, 2007.

[86] J. Bajorath, Ed., *Chemoinformatics and Computational Chemical Biology*, ser. Springer Protocols 672. New York, NY: Humana Press, 2011, ISBN: 978-1-60761-838-6.

[87] T. Engel and J. Gasteiger, *Chemoinformatics: Basic Concepts and Methods*. John Wiley & Sons, 2018.

[88] T. Engel, "Basic Overview of Chemoinformatics," *Journal of Chemical Information and Modeling*, vol. 46, no. 6, pp. 2267–2277, Nov. 2006, ISSN: 1549-9596. DOI: 10.1021/ci600234z.

[89] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, "DrugBank: A comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D668–D672, Jan. 2006, ISSN: 0305-1048. DOI: 10.1093/nar/gkj067.

[90] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, “DrugBank: A knowledgebase for drugs, drug actions and drug targets,” *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D901–D906, Jan. 2008, ISSN: 0305-1048. DOI: 10.1093/nar/gkm958.

[91] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, “DrugBank 3.0: A comprehensive resource for ‘Omics’ research on drugs,” *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D1035–D1041, Jan. 2011, ISSN: 0305-1048. DOI: 10.1093/nar/gkq1126.

[92] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, “DrugBank 4.0: Shedding new light on drug metabolism,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D1091–D1097, Jan. 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gkt1068.

[93] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, “DrugBank 5.0: A major update to the DrugBank database for 2018,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, ISSN: 0305-1048. DOI: 10.1093/nar/gkx1037.

[94] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, “ChEMBL: A large-scale bioactivity database for drug discovery,” *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, Jan. 2012, ISSN: 0305-1048. DOI: 10.1093/nar/gkr777.

[95] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington, “The ChEMBL bioactivity database: An update,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D1083–D1090, Jan. 2014, ISSN: 0305-1048. DOI: 10.1093/nar/gkt1031.

[96] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach, “The ChEMBL database in 2017,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, ISSN: 0305-1048. DOI: 10.1093/nar/gkw1074.

[97] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, “The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data,” *Nucleic Acids Research*, vol. 35, no. Database, pp. D301–D303, Jan. 2007, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkl971.

[98] wwPDB consortium, "Protein Data Bank: The single global archive for 3D macromolecular structure data," *Nucleic Acids Research*, vol. 47, no. D1, pp. D520–D528, Jan. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gky949.

[99] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, Jan. 2000, ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.

[100] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, ISSN: 1549-9596. DOI: 10.1021/ci00057a005.

[101] W. T. Wipke and T. M. Dyott, "Stereochemically unique naming algorithm," *Journal of the American Chemical Society*, vol. 96, no. 15, pp. 4834–4842, Jul. 1974, ISSN: 0002-7863, 1520-5126. DOI: 10.1021/ja00822a021.

[102] N. Schneider, R. A. Sayle, and G. A. Landrum, "Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm," *Journal of Chemical Information and Modeling*, vol. 55, no. 10, pp. 2111–2120, Oct. 2015, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.5b00543.

[103] *Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description*, Version 3.30. wwPDB, Nov. 2012.

[104] "Tripos Mol2 File Format," Tech. Rep. [Online]. Available: http://chemyang.ccnu.edu.cn/ccb/server/AIMMS/mol2.pdf (visited on 10/02/2018).

[105] P. C. D. Hawkins, "Conformation Generation: The State of the Art," *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 1747–1756, Aug. 2017, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.7b00221.

[106] J.-P. Ebejer, G. M. Morris, and C. M. Deane, "Freely Available Conformer Generation Methods: How Good Are They?" *Journal of Chemical Information and Modeling*, vol. 52, no. 5, pp. 1146–1158, May 2012, ISSN: 1549-9596. DOI: 10.1021/ci2004658.

[107] P. C. D. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, and M. T. Stahl, "Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database," *Journal of Chemical Information and Modeling*, vol. 50, no. 4, pp. 572–584, Apr. 2010, ISSN: 1549-9596. DOI: 10.1021/ci100031x.

[108] J. Boström, “Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools,” *Journal of Computer-Aided Molecular Design*, vol. 15, no. 12, pp. 1137–1152, Dec. 2001, ISSN: 1573-4951. DOI: 10.1023/A:1015930826903.

[109] J. Boström, J. R. Greenwood, and J. Gottfries, “Assessing the performance of OMEGA with respect to retrieving bioactive conformations,” *Journal of Molecular Graphics and Modelling*, vol. 21, no. 5, pp. 449–462, Mar. 2003, ISSN: 1093-3263. DOI: 10.1016/S1093-3263(02)00204-8.

[110] E. E. Bolton, S. Kim, and S. H. Bryant, “PubChem3D: Conformer generation,” *Journal of Cheminformatics*, vol. 3, no. 1, p. 4, Jan. 2011, ISSN: 1758-2946. DOI: 10.1186/1758-2946-3-4.

[111] P. C. D. Hawkins and A. Nicholls, “Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures,” *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2919–2936, Nov. 2012, ISSN: 1549-9596. DOI: 10.1021/ci300314k.

[112] T. A. Halgren, “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94,” *Journal of Computational Chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996, ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(199604)17:5/6⟨490::AID-JCC1⟩3.0.CO;2-P.

[113] *OMEGA*, OpenEye Scientific Software, Santa Fe, NM. [Online]. Available: http://www.eyesopen.com.

[114] N. Nikolova and J. Jaworska, “Approaches to Measure Chemical Similarity – a Review,” *QSAR & Combinatorial Science*, vol. 22, no. 9-10, pp. 1006–1026, 2003, ISSN: 1611-0218. DOI: 10.1002/qsar.200330831.

[115] A. G. Maldonado, J. P. Doucet, M. Petitjean, and B.-T. Fan, “Molecular similarity and diversity in chemoinformatics: From theory to applications,” *Molecular Diversity*, vol. 10, no. 1, pp. 39–79, Feb. 2006, ISSN: 1573-501X. DOI: 10.1007/s11030-006-8697-1.

[116] P. Willett, “Similarity Methods in Chemoinformatics,” *Annual Review of Information Science and Technology*, no. 43, pp. 3–71, 2009.

[117] D. Stumpfe and J. Bajorath, “Similarity searching,” *WIREs Computational Molecular Science*, vol. 1, no. 2, pp. 260–282, 2011, ISSN: 1759-0884. DOI: 10.1002/wcms.23.

[118] M. Vogt and J. Bajorath, “Modeling Tanimoto Similarity Value Distributions and Predicting Search Results,” *Molecular Informatics*, vol. 36, no. 7, p. 1 600 131, 2017, ISSN: 1868-1751. DOI: 10.1002/minf.201600131.

[119] G. M. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, “Molecular Similarity in Medicinal Chemistry,” *Journal of Medicinal Chemistry*, vol. 57, no. 8, pp. 3186–3204, Apr. 2014, ISSN: 0022-2623. DOI: 10.1021/jm401411z.

[120] P. Ripphausen, B. Nisius, and J. Bajorath, “State-of-the-art in ligand-based virtual screening,” *Drug Discovery Today*, vol. 16, no. 9, pp. 372–376, May 2011, ISSN: 1359-6446. DOI: 10.1016/j.drudis.2011.02.011.

[121] G. M. Maggiora and M. A. Johnson, Eds., *Concepts and Applications of Molecular Similarity*. Wiley, Sep. 1990, ISBN: 978-0-471-62175-1. [Online]. Available: https://www.wiley.com/en-us/Concepts+and+Applications+of+Molecular+Similarity-p-9780471621751.

[122] F. Barbosa and D. Horvath, “Molecular Similarity and Property Similarity,” *Current Topics in Medicinal Chemistry*, vol. 4, no. 6, pp. 589–600, Feb. 2004, ISSN: 15680266. DOI: 10.2174/1568026043451186.

[123] G. M. Maggiora, “On Outliers and Activity Cliffs - Why QSAR Often Disappoints,” *Journal of Chemical Information and Modeling*, vol. 46, no. 4, pp. 1535–1535, Jul. 2006, ISSN: 1549-9596. DOI: 10.1021/ci060117s.

[124] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, “Molecular fingerprint similarity search in virtual screening,” *Methods*, Virtual Screening, vol. 71, pp. 58–63, Jan. 2015, ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2014.08.005.

[125] D. Bajusz, A. Rácz, and K. Héberger, “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?” *Journal of Cheminformatics*, vol. 7, no. 1, p. 20, May 2015, ISSN: 1758-2946. DOI: 10.1186/s13321-015-0069-3.

[126] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, “The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics,” *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 2, pp. 493–500, Mar. 2003, ISSN: 0095-2338. DOI: 10.1021/ci025584y.

[127] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck, “The Chemistry Development Kit (CDK) v2.0: Atom typing, depiction, molecular formulas, and substructure searching,” *Journal of Cheminformatics*, vol. 9, no. 1, p. 33, Jun. 2017, ISSN: 1758-2946. DOI: 10.1186/s13321-017-0220-4.

[128] E. Willighagen, *Groovy Cheminformatics with the Chemistry Development Kit*, 2.3.2. EL Willighagen, 2011. [Online]. Available: https://egonw.github.io/cdkbook/.

[129] *Daylight Theory Manual.* Aug. 2011. [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/.

[130] J. A. Haigh, B. T. Pickup, J. A. Grant, and A. Nicholls, "Small Molecule Shape-Fingerprints," *Journal of Chemical Information and Modeling*, vol. 45, no. 3, pp. 673–684, May 2005, ISSN: 1549-9596. DOI: 10.1021/ci049651v.

[131] J. A. Grant, M. A. Gallardo, and B. T. Pickup, "A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape," *Journal of Computational Chemistry*, vol. 17, no. 14, pp. 1653–1666, 1996, ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(19961115)17:14⟨1653::AID-JCC7⟩3.0.CO;2-K.

[132] J. A. Grant and B. T. Pickup, "A Gaussian Description of Molecular Shape," *The Journal of Physical Chemistry*, vol. 99, no. 11, pp. 3503–3510, Mar. 1995, ISSN: 0022-3654. DOI: 10.1021/j100011a016.

[133] *ROCS*, OpenEye Scientific Software, Santa Fe, NM. [Online]. Available: http://www.eyesopen.com.

[134] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha, "QSAR Modeling: Where Have You Been? where Are You Going To?" *Journal of Medicinal Chemistry*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, ISSN: 0022-2623. DOI: 10.1021/jm4004285.

[135] S. Yousefinejad and B. Hemmateenejad, "Chemometrics tools in QSAR/QSPR studies: A historical perspective," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 177–204, Dec. 2015, ISSN: 0169-7439. DOI: 10.1016/j.chemolab.2015.06.016.

[136] A. Varnek and I. Baskin, "Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis?" *Journal of Chemical Information and Modeling*, vol. 52, no. 6, pp. 1413–1437, Jun. 2012, ISSN: 1549-9596, 1549-960X. DOI: 10.1021/ci200409x.

[137] P. Gramatica, "Principles of QSAR models validation: Internal and external," *QSAR & Combinatorial Science*, vol. 26, no. 5, pp. 694–701, 2007, ISSN: 1611-0218. DOI: 10.1002/qsar.200610151.

[138] A. Burkov, *The Hundred-Page Machine Learning Book.* 2019, ISBN: 978-1-9995795-0-0 978-1-9995795-1-7. [Online]. Available: http://themlbook.com.

[139] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine Learning - Algorithms and Applications.* CRC Press, 2017.

[140] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014, ISBN: 978-1-107-29801-9. DOI: 10.1017/CBO9781107298019.

[141] G. Bonaccorso, *Machine Learning Algorithms*. Birmingham Mumbai: Packt, 2017, ISBN: 978-1-78588-962-2.

[142] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

[143] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference and Prediction*, Second. Springer, 2008.

[144] P. Franco, "Orphan drugs: The regulatory environment," *Drug Discovery Today*, vol. 18, no. 3, pp. 163–172, Feb. 2013, ISSN: 1359-6446. DOI: 10.1016/j.drudis.2012.08.009.

[145] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *Journal of Health Economics*, vol. 47, pp. 20–33, May 2016, ISSN: 0167-6296. DOI: 10.1016/j.jhealeco.2016.01.012.

[146] S. Morgan, P. Grootendorst, J. Lexchin, C. Cunningham, and D. Greyson, "The cost of drug development: A systematic review," *Health Policy*, vol. 100, no. 1, pp. 4–17, Apr. 2011, ISSN: 0168-8510. DOI: 10.1016/j.healthpol.2010.12.002.

[147] S. Simmons and Z. Estes, "Individual differences in the perception of similarity and difference," *Cognition*, vol. 108, no. 3, pp. 781–795, Sep. 2008, ISSN: 0010-0277. DOI: 10.1016/j.cognition.2008.07.003.

[148] P. S. Kutchukian, N. Y. Vasilyeva, J. Xu, M. K. Lindvall, M. P. Dillon, M. Glick, J. D. Coley, and N. Brooijmans, "Inside the Mind of a Medicinal Chemist: The Role of Human Bias in Compound Prioritization during Drug Discovery," *PLOS ONE*, vol. 7, no. 11, e48476, Nov. 2012, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0048476.

[149] M. S. Lajiness, G. M. Maggiora, and V. Shanmugasundaram, "Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds," *Journal of Medicinal Chemistry*, vol. 47, no. 20, pp. 4891–4896, Sep. 2004, ISSN: 0022-2623. DOI: 10.1021/jm049740z.

[150] M. D. Hack, D. N. Rassokhin, C. Buyck, M. Seierstad, A. Skalkin, P. ten Holte, T. K. Jones, T. Mirzadegan, and D. K. Agrafiotis, "Library Enhancement through the Wisdom of Crowds," *Journal of Chemical Information and Modeling*, vol. 51, no. 12, pp. 3275–3286, Dec. 2011, ISSN: 1549-9596. DOI: 10.1021/ci200446y.

[151] P. Franco, N. Porta, J. D. Holliday, and P. Willett, "The use of 2D fingerprint methods to support the assessment of structural similarity in orphan drug legislation," *Journal of Cheminformatics*, vol. 6, no. 1, p. 5, Feb. 2014, ISSN: 1758-2946. DOI: 10.1186/1758-2946-6-5.

[152] P. Franco, N. Porta, J. D. Holliday, and P. Willett, "Molecular similarity considerations in the licensing of orphan drugs," *Drug Discovery Today*, vol. 22, no. 2, pp. 377–381, Feb. 2017, ISSN: 1359-6446. DOI: 10.1016/j.drudis.2016.11.024.

[153] *Molecular Operating Environment*, Chemical Computing Group ULC, 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2020.

[154] P. C. D. Hawkins, A. G. Skillman, and A. Nicholls, "Comparison of Shape-Matching and Docking as Virtual Screening Tools," *Journal of Medicinal Chemistry*, vol. 50, no. 1, pp. 74–82, Jan. 2007, ISSN: 0022-2623. DOI: 10.1021/jm0603365.

[155] G. Cruciani, P. Crivori, P.-A. Carrupt, and B. Testa, "Molecular fields in quantitative structure–permeation relationships: The VolSurf approach," *Journal of Molecular Structure: THEOCHEM*, vol. 503, no. 1, pp. 17–30, May 2000, ISSN: 0166-1280. DOI: 10.1016/S0166-1280(99)00360-7.

[156] G. Cruciani, M. Pastor, and W. Guba, "VolSurf: A new tool for the pharmacokinetic optimization of lead compounds," *European Journal of Pharmaceutical Sciences*, Frontiers in Biopharmacy, vol. 11, S29–S39, Oct. 2000, ISSN: 0928-0987. DOI: 10.1016/S0928-0987(00)00162-7.

[157] A. Artese, S. Cross, G. Costa, S. Distinto, L. Parrotta, S. Alcaro, F. Ortuso, and G. Cruciani, "Molecular interaction fields in drug discovery: Recent advances and future perspectives," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, no. 6, pp. 594–613, 2013, ISSN: 1759-0884. DOI: 10.1002/wcms.1150.

[158] G. Cruciani, R. Mannhold, H. Kubinyi, and G. Folkers, Eds., *Molecular Interaction Fields - Applications in Drug Discovery and ADME Prediction*, ser. Methods and Principles in Medicinal Chemistry 27. Wiley-VCH Verlag GmbH, 2006.

[159] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, May 2010, ISSN: 1549-9596. DOI: 10.1021/ci100050t.

[160] *RDKit: Open-source cheminformatics*, Apr. 2013. [Online]. Available: http://www.rdkit.org.

[161] M. Swain, *MolVS: Molecule Validation and Standardization*, Aug. 2021. [Online]. Available: https://github.com/mcs07/MolVS (visited on 08/18/2021).

[162] M. Claesen and B. De Moor, "Hyperparameter Search in Machine Learning," *arXiv:1502.02127 [cs, stat]*, Apr. 2015. arXiv: 1502.02127 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1502.02127 (visited on 08/25/2021).

[163] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, ISSN: 2231007X, 22309608. DOI: 10.5121/ijdkp.2015.5201.

[164] M. C. Sanguinetti and M. Tristani-Firouzi, "hERG potassium channels and cardiac arrhythmia," *Nature*, vol. 440, no. 7083, pp. 463–469, Mar. 2006, ISSN: 1476-4687. DOI: 10.1038/nature04710.

[165] B. L. Roth, "Drugs and Valvular Heart Disease," *New England Journal of Medicine*, vol. 356, no. 1, pp. 6–9, Jan. 2007, ISSN: 0028-4793. DOI: 10.1056/NEJMp068265.

[166] B. Wang, L.-P. Yang, X.-Z. Zhang, S.-Q. Huang, M. Bartlam, and S.-F. Zhou, "New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme," *Drug Metabolism Reviews*, vol. 41, no. 4, pp. 573–643, Nov. 2009, ISSN: 0360-2532. DOI: 10.1080/03602530903118729.

[167] K. Wilson and J. M. Walker, Eds., *Principles and Techniques of Biochemistry and Molecular Biology*, 7th ed. Cambridge, UK : New York: Cambridge University Press, 2009, ISBN: 978-0-521-51635-8 978-0-521-73167-6.

[168] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, "InChI - the worldwide chemical structure identifier standard," *Journal of Cheminformatics*, vol. 5, no. 1, p. 7, Jan. 2013, ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-7.

[169] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi, "InChI, the IUPAC International Chemical Identifier," *Journal of Cheminformatics*, vol. 7, no. 1, p. 23, May 2015, ISSN: 1758-2946. DOI: 10.1186/s13321-015-0068-4.

[170] D. B. Rorabacher, "Statistical treatment for rejection of deviant values: Critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level," *Analytical Chemistry*, vol. 63, no. 2, pp. 139–146, Jan. 1991, ISSN: 0003-2700, 1520-6882. DOI: 10.1021/ac00002a010.

[171] J. N. Ehrman, V. T. Lim, C. C. Bannan, N. Thi, D. Y. Kyu, and D. L. Mobley, "Improving small molecule force fields by identifying and characterizing small molecules with inconsistent parameters," *Journal of Computer-Aided Molecular Design*, Jan. 2021, ISSN: 1573-4951. DOI: 10.1007/s10822-020-00367-1.

[172] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature Chemistry*, vol. 4, no. 2, pp. 90–98, Feb. 2012, ISSN: 1755-4349. DOI: 10.1038/nchem.1243.

[173] A. Nicholls, G. B. McGaughey, R. P. Sheridan, A. C. Good, G. Warren, M. Mathieu, S. W. Muchmore, S. P. Brown, J. A. Grant, J. A. Haigh, N. Nevins, A. N. Jain, and B. Kelley, "Molecular shape and medicinal chemistry: A perspective.," *Journal of Medicinal Chemistry*, vol. 53, no. 10, pp. 3862–3886, May 2010, ISSN: 0022-2623, 1520-4804. DOI: 10.1021/jm900818s.

[174] L. C. Blum, R. van Deursen, and J.-L. Reymond, "Visualisation and subsets of the chemical universe database GDB-13 for virtual screening," *Journal of Computer-Aided Molecular Design*, vol. 25, no. 7, pp. 637–647, Jul. 2011, ISSN: 1573-4951. DOI: 10.1007/s10822-011-9436-y.

[175] T. Sander, J. Freyss, M. von Korff, and C. Rufener, "DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis," *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 460–473, Feb. 2015, ISSN: 1549-9596. DOI: 10.1021/ci500588j.

[176] P. Franco, "Similarity in the context of the Orphan Drug egislation Legislation," PhD thesis, University of Sheffield, 2015.

[177] A. S. Rose and P. W. Hildebrand, "NGL Viewer: A web application for molecular visualization," *Nucleic Acids Research*, vol. 43, no. W1, W576–W579, Jul. 2015, ISSN: 0305-1048. DOI: 10.1093/nar/gkv402.

[178] H. Nguyen, D. A. Case, and A. S. Rose, "NGLview–interactive molecular graphics for Jupyter notebooks," *Bioinformatics*, vol. 34, no. 7, A. Valencia, Ed., pp. 1241–1242, Apr. 2018, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx789.

[179] *Voila-dashboards/voila*, Voilà Dashboards, Aug. 2021. [Online]. Available: https://github.com/voila-dashboards/voila (visited on 08/22/2021).

[180] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter Notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds., IOS Press, 2016, pp. 87–90. DOI: 10.3233/978-1-61499-649-1-87.

[181] *Heroku - Cloud Application Platform*. [Online]. Available: https://www.heroku.com/ (visited on 08/22/2021).

[182] P. G. D. Group, *PostgreSQL*, 2021-08-22T15:32:16.585153. [Online]. Available: https://www.postgresql.org/ (visited on 08/22/2021).

[183] M. Bayer, “SQLAlchemy,” in *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*, A. Brown and G. Wilson, Eds., Mountain View: aosabook.org, 2012. [Online]. Available: `http://aosabook.org/en/sqlalchemy.html`.

[184] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” p. 25,

[185] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*, ser. The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019, ISBN: 978-3-030-05317-8 978-3-030-05318-5. DOI: `10.1007/978-3-030-05318-5`.