Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Avogadro–Scaled Hypergeometric–Chebyshev Feature Maps and Entanglement-Inspired Kernels for Generative-Docking Pattern Recognition in SARS-CoV-2 Targets

IOANNIS GRIGORIADIS

JGRIGORIADIS@BIOGENEA.GR

Biogenea Pharmaceuticals (Greece)

Research Article

Keywords: molecular docking, kernel methods, feature maps, generative ligand design, SARS-CoV-2, virtual screening, Pattern recognition, molecular representation learning, Chebyshev polynomials, hypergeometric functions, generative-docking

Posted Date: April 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9361450/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Avogadro–Scaled Hypergeometric–Chebyshev Feature Maps and Entanglement-Inspired Kernels for Generative-Docking Pattern Recognition in SARS-CoV-2 Targets

Ioannis Grigoriadis (I.G.)

Biogenea™ — Department of Computer Drug Discovery Science (Biogenetoligandorol™ Stations), Thessaloniki, Greece

QAvogadro™ Stations, Thessaloniki, Greece

Roccuffirna™ R&D Units, Thessaloniki, Greece

Biogenea Pharmaceuticals Ltd, Thessaloniki, Greece

Correspondence: jgrigoriadis@biogenea.gr | +30 6936592686

ORCID: 0000-0002-8443-1777

Note on trademark/IP: "Rocuffirna™" is used as a working trademark / candidate-series name in this manuscript (see Declaration of Interest).

Abstract

We describe a docking-oriented pattern-recognition framework for structure-based ligand generation and screening that combines (i) Avogadro-number scaling as a deterministic magnitude stabilizer and (ii) special-function feature maps—regularized hypergeometric expansions and Chebyshev polynomial embeddings—as high-capacity descriptors of pocket–ligand interactions. Similarity is implemented via an entanglement-inspired positive semidefinite kernel that behaves like an operator overlap (trace) after partial-trace-like aggregation over fragment ensembles. Candidate ligands are assembled by a rule-driven generator that treats fragment assembly as a restricted formal language, enforcing syntactic chemical validity by construction. The method is evaluated strictly as an in-silico docking-oriented pattern-recognition system across multiple SARS-CoV-2 targets (spike-related and protease-related binding pockets) using generated ligand families compared to reference antivirals. Outputs are reported as patterns over docking affinity and energy components (van der Waals, electrostatics, total), emphasizing numerical stability, rank behavior, and failure modes. No clinical efficacy is claimed; results remain computational hypotheses requiring experimental validation.

Keywords: Pattern recognition; molecular representation learning; Chebyshev polynomials; hypergeometric functions; kernel methods; generative-docking; SARS-CoV-2.

---

1. Introduction

Docking-oriented molecular pattern recognition treats docking as an **evaluation oracle** rather than the primary optimization object: under a fixed protocol, docking returns a reproducible scoreboard $D(P, L)$ for pocket–ligand pairs, optionally accompanied by component terms such as $(E_{\mathrm{vdW}}, E_{\mathrm{elec}}, E_{\mathrm{tot}})$, while the learning problem is shifted to the representation, similarity, and constrained-generation stack that makes this scoreboard stable and usable at scale [1]–[4], [52]–[55]. This framing is motivated by two recurrent failure modes in practical screening: first, pocket–ligand structure can be encoded in ways that collapse under multiscale dynamic range, truncation, or poor conditioning; second, generative search can yield invalid chemistry or low-diversity candidates before docking capacity is meaningfully exploited [1]–[8], [23]–[26]. Accordingly, the method is designed around two linked requirements—representational expressivity and downstream rank alignment—summarized operationally as **"Scale → Expand → Translate → Assemble → Aggregate → Score → Compare"**. In the present study, "Scale" denotes deterministic magnitude conditioning such as $\tilde{u} = u/(N_A^{\alpha} + \varepsilon)$ **(1)**, "Expand" denotes bounded nonlinear lifting such as $\hat{u} = \tanh(\tilde{u})$ **(2)** followed by special-function embeddings, and "Aggregate/Compare" denotes PSD pooling and overlap computation, e.g. $K(P, L) = \mathrm{Tr}(\rho_P \rho_L)$ **(3)**, so that the final score can be interpreted as a stable, rank-oriented surrogate of the docking oracle rather than as an absolute physical observable [23]–[26], [30]–[36]. The application context is entirely **in silico** and targets SARS-CoV-2 spike-related complexes and the main protease (Mpro) derived from experimentally determined Protein Data Bank structures [9]–[13]. All outputs are therefore computational hypotheses: docking does not establish efficacy or safety, and meaningful translation still requires synthesis feasibility, biochemical testing, and broader pharmacological profiling [14], [15], [52]–[55]. This scope boundary is important given the broader COVID-19 treatment landscape, which extends well beyond any single computational workflow [16], [52].

2. Methods and materials: representation, ranking, and constrained generation

2.1 Rank-aligned docking-oriented representation and PSD kernel scoring pipeline

The method is designed around the premise that, in docking-oriented generative screening, the main challenge is not evaluating a fixed docking score $D(P, L)$ but constructing ligand representations that remain chemically valid, diverse, pocket-aware, and comparable under a stable similarity notion [17]–[26], [31]–[36]. Under a fixed docking protocol, each pocket–ligand pair is assigned an oracle score $D(P, L)$, and pocket-specific preferences are defined by $L_i \succ L_j \Leftrightarrow D(P, L_i) < D(P, L_j)$ **(1)**, so the learning problem is cast as ranking rather than absolute score prediction [1]–[4], [27]–[29], [52]–[55]. To stabilize features before nonlinear expansion, each channel u is conditioned as $\tilde{u} = u/(N_A^{\alpha} + \varepsilon)$ **(2)** and then bounded by $\hat{u} = \tanh(\tilde{u})$ **(3)**, after which pocket channels are expanded by Chebyshev carriers and ligand channels by truncated regularized hypergeometric maps [30]–[36]. Fragment and pocket embeddings are pooled into PSD operators $\rho_L = \sum_m w_m \ \phi(f_m)\phi(f_m)^{\top}$ **(4)** and $\rho_P = \sum_r \eta_r \ \psi(p_r)\psi(p_r)^{\top}$ **(5)**, and compatibility is measured

by the trace-overlap kernel $K(P,L) = \mathrm{Tr}(\rho_P \rho_L)$ **(6)**, optionally regularized by shrinkage $\rho \leftarrow (1-\tau)\rho + \tau I/d$ **(7)** to improve conditioning while preserving PSD structure [23]–[26], [37]–[39]. The final score is formed by fusing explicit interaction descriptors and kernel diagnostics, e.g. $\Psi(P,L) = [\Phi(P,L) \,\|\, K(P,L) \,\|\, \log(K(P,L)+\delta) \,\|\, \mathrm{Tr}(\rho_P) \,\|\, \mathrm{Tr}(\rho_L)]$ **(8)**, followed by either a linear predictor $S_\theta(P,L) = \beta^\top \Psi(P,L)$ **(9)** or a nonlinear map $S_\theta(P,L) = \mathrm{MLP}_\theta(\Psi(P,L))$ **(10)** [23]–[26]. Rank alignment is then learned with a pairwise logistic objective $L(\theta;P) = \sum_{(i,j)\in\mathcal{P}(P)} \log \sigma(1 + \exp(m - [S_\theta(P,L_i) - S_\theta(P,L_j)]))$ **(11)**, optionally weighted by docking hardness $w_{ij} = 1/(|\, D(P,L_i) - D(P,L_j)\,| + \varepsilon)$ **(12)**, reflecting the emphasis on enrichment and rank agreement in docking benchmarks [27]–[29].

2.2 Constrained fragment generation, diversity-controlled beam search, and experimental screening setup

Ligand generation is treated as constrained symbolic execution rather than unconstrained string sampling, so that validity is enforced during construction rather than repaired afterward [5]–[8], [40]–[49]. A ligand is generated from a fragment library F and grammar G as $L \sim G$ subject to $C(L) = \mathrm{true}$ **(13)**, where $C(\cdot)$ encodes valence, bond-compatibility, ring-closure, attachment-type, and optional stereochemical constraints; operationally, an action a_t is accepted only if $\mathrm{Valid}(s_t, a_t) = 1$ **(14)** and the state is updated by $(s_{t+1}, L_{t+1}) = \mathrm{Step}(s_t, L_t, a_t)$ **(15)** [40], [41]. Search is performed by a kernel-guided beam in which the partial objective is $\mathrm{Score}(P,L) = S_\theta(P,L) + \gamma h(L)$ **(16)**, for example with a length penalty $h_{\mathrm{len}}(L) = -\alpha\,|\,L\,|$ **(17)**, and diversity is protected by a fingerprint penalty $\mathrm{DivPen}(L;B) = \max_{L'\in B} \mathrm{Sim}_{\tan}(L,L')$ **(18)** using $\mathrm{Sim}_{\tan}(L,L') = \frac{\langle f(L), f(L')\rangle}{\|f(L)\|_1 + \|f(L')\|_1 - \langle f(L), f(L')\rangle}$ **(19)**, so that $\mathrm{Score}_{\mathrm{div}}(P,L) = \mathrm{Score}(P,L) - \eta\,\mathrm{DivPen}(L;B)$ **(20)** and $B_{t+1} = \mathrm{TopB}(C_{t+1}; \mathrm{Score}_{\mathrm{div}}(P,\cdot))$ **(21)** [17]–[21]. Experimentally, targets are selected from SARS-CoV-2 spike-related and protease-related PDB structures [9]–[13], candidate pockets are defined by cavity detection such as $P(T) = \{P_1, \ldots, P_{n_P(T)}\}$ **(22)** using fpocket-like procedures [50], and each generated ligand is conformer-prepared and minimized before docking under a fully logged protocol because docking rankings remain sensitive to pocket definition, ligand preparation, and engine settings [1]–[4], [19], [50]–[55].

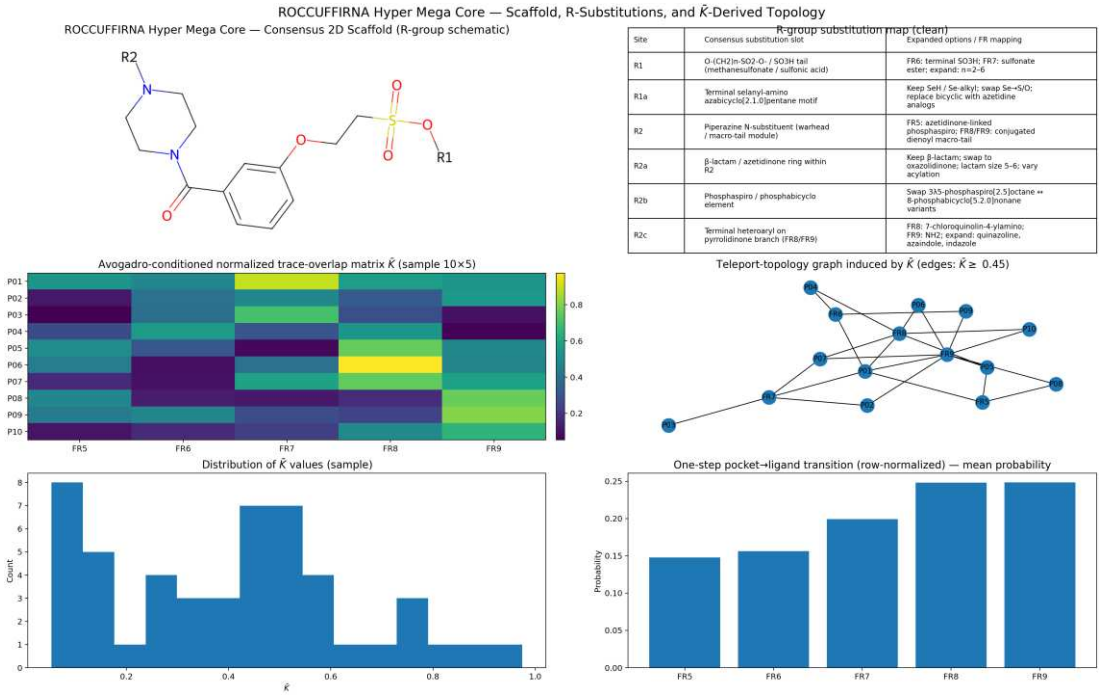


Figure 1. **ROCCUFFIRNA Hyper Mega Core scaffold, R-substitution grammar, and $\bar{K}$-derived topology under Avogadro-conditioned operator kernels. (A)** Consensus 2D scaffold of the ROCCUFFIRNA Hyper Mega Core showing the two design handles: **R1** (sulfonate O-substituent / acid–ester tail) and **R2** (distal piperazine N-substituent / warhead–macro-tail module). **(B)** Publication-facing "substitution map" that links each handle (and sub-sites) to the FR5–FR9 variant logic (e.g., R1 capturing terminal $-SO_3H$ vs sulfonate-ester tails; R2 capturing β-lactam/phosphaspiro and heteroaryl macro-tail motifs). **(C)** Avogadro-conditioned **normalized trace-overlap** matrix $\bar{K}$ (sample 10×5) comparing pocket operators $\{\rho_{P_i}\}_{i=1}^{10}$ (rows P01–P10) against the fixed ligand panel $\{\rho_{FR_j}\}_{j\in\{5,\ldots,9\}}$ (columns FR5–FR9). Operators are constructed as PSD covariances of expanded, bounded features (Chebyshev for pockets; truncated hypergeometric for ligands) after robust normalization and deterministic Avogadro scaling: $z_k = \frac{f_k - \mathrm{median}(f_k)}{\mathrm{MAD}(f_k)+\varepsilon}, s(\alpha) = \exp(-\alpha \log N_A), \hat{f} = \tanh\left(z \cdot s(\alpha)\right), \rho = \frac{1}{n} Z^\top Z \succeq 0, K(P_i, FR_j) = \mathrm{Tr}\left(\rho_{P_i}\rho_{FR_j}\right), \bar{K}(P_i, FR_j) = \frac{\mathrm{Tr}\left(\rho_{P_i}\rho_{FR_j}\right)}{\sqrt{\mathrm{Tr}\left(\rho_{P_i}^2\right)\mathrm{Tr}\left(\rho_{FR_j}^2\right)}+\varepsilon}$. The color scale highlights heterogeneous pocket–ligand compatibility patterns (higher $\bar{K} \Rightarrow$ stronger operator overlap under the PSD kernel). **(D)** Empirical distribution of $\bar{K}$ values across the sampled pocket–ligand pairs, summarizing similarity mass and spread (useful for diagnosing saturation/collapse vs discriminative separation under the chosen expansion orders and shrinkage). **(E)** "Teleport-topology" graph induced by thresholding $\bar{K}$: an edge is drawn when $\bar{K} \geq 0.45$, yielding a sparse connectivity view that can be used as the substrate for teleportation/quantum-walk style margin enhancement (graph construction is strictly representation-derived, not a physical claim). **(F)** One-step pocket→ligand transition probabilities obtained by row-normalizing $\bar{K}$, $p_{i\to j} = \frac{\bar{K}_{ij}}{\sum_{j'} \bar{K}_{ij'}+\varepsilon}$, and then averaging across pockets (mean probabilities shown): FR5 ≈ 0.15, FR6 ≈ 0.16, FR7 ≈ 0.20, FR8 ≈ 0.25, FR9 ≈ 0.25. This aggregation provides a compact, reproducible summary of which FR variants dominate the induced affinity flow under the Avogadro-conditioned normalized trace-overlap kernel.

2.2.1 Conformer ensembles and minimization

For each ligand L, a conformer ensemble is generated using RDKit [19], $\mathrm{Conf}(L) = \{\chi_1, \dots, \chi_C\}$**(1)**, and each conformer is then minimized under MMFF94 [51] as $\chi_c^\star = \arg\min_{\chi_c} E_{\mathrm{MMFF94}}(\chi_c)$, $c = 1, \dots, C$ **(2)**, to reduce strain and improve consistency before docking, which is important because docking outcomes are sensitive to ligand preparation and starting geometry [1]–[4], [52]–[55]. Per-ligand summary statistics are reported as the mean minimized conformer energy $\bar{E}(L) = \frac{1}{C}\sum_{c=1}^{C} E_{\mathrm{MMFF94}}(\chi_c^\star)$ **(3)** and the minimum minimized conformer energy $E_{\min}(L) = \min_c E_{\mathrm{MMFF94}}(\chi_c^\star)$ **(4)**; these quantities can be used as heuristic terms $h(L)$ in beam search and as filters for extreme conformational strain [19], [51].

2.2.2 Fingerprints and graph descriptors for baselines and diversity

Each ligand is also encoded by a Morgan/ECFP-like fingerprint $f(L)$ [17], [18], which supports baseline similarity scoring, diversity control, and nearest-neighbor analysis against reference antivirals. In particular, pairwise Tanimoto similarity is computed as $\mathrm{Sim}_{\tan}(L_i, L_j) = \frac{\langle f(L_i), f(L_j)\rangle}{\|f(L_i)\|_1 + \|f(L_j)\|_1 - \langle f(L_i), f(L_j)\rangle}$ **(5)**, enabling a fingerprint baseline $S_{\mathrm{fp}}(L) = \max_{r \in R} \mathrm{Sim}_{\tan}(L, r)$ **(6)**over a reference set R, as well as diversity tracking through $\mathrm{Div} = \frac{2}{N(N-1)} \sum_{i<j} (1 - \mathrm{Sim}_{\tan}(L_i, L_j))$ **(7)** [17]–[21]. This use of ECFP-style encodings follows standard cheminformatics practice and provides a compact baseline against which the more structured docking- and kernel-based representations can be assessed [17]–[21].

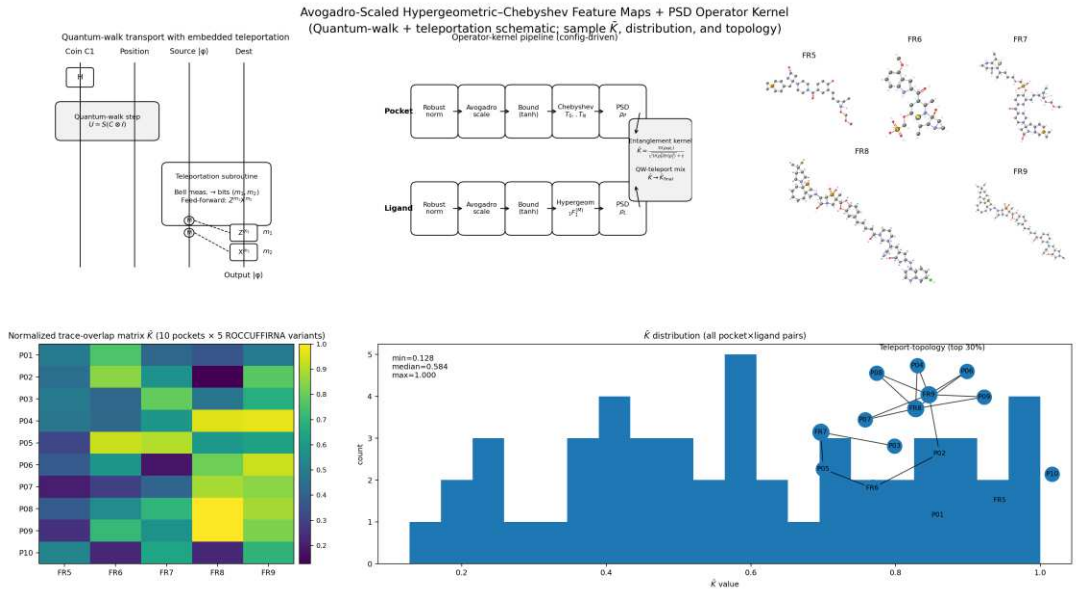


Figure 2. ***Avogadro-scaled hypergeometric–Chebyshev feature maps with PSD operator overlap, and teleportation/quantum-walk margin enhancement for ROCCUFFIRNA FR5–FR9 pocket–ligand pattern recognition.*** *The composite figure summarizes a config-driven pipeline that compares a collection of protein pockets* $\{P_i\}_{i=1}^{N}$ *against a fixed ROCCUFFIRNA ligand panel* $\mathcal{L} = \{FR5, FR6, FR7, FR8, FR9\}$ *by mapping pocket and ligand point-cloud descriptors into positive semidefinite (PSD) operators and*

evaluating their overlap. Protein structures S_i are ingested and converted into pocket regions P_i, from which pocket point sets $X_i = \{x_{i,n}\}_{n=1}^{n_i} \subset \mathbb{R}^3$ are sampled; ligands are parsed into ligand point sets $Y_j = \{y_{j,m}\}_{m=1}^{m_j} \subset \mathbb{R}^3$ for each $L_j \in \mathcal{L}$. Multi-channel physicochemical features $\phi_P(x) \in \mathbb{R}^{d_P}$ and $\phi_L(y) \in \mathbb{R}^{d_L}$ are stabilized by robust channel normalization and Avogadro conditioning prior to series expansion, using channelwise median/MAD normalization $z_k = (f_k - median(f_k))/(MAD(f_k) + \varepsilon)$ followed by a deterministic magnitude conditioner anchored at Avogadro's constant $N_A = 6.02214076 \times 10^{23}$ applied stably in log-space as $s(\alpha) = exp\,(-\alpha log\ N_A)$ so that $\tilde{f} = z\,s(\alpha)$; bounded inputs for polynomial bases are then enforced by $\hat{f} = tanh\,(\tilde{f}) \in (-1,1)$. Pocket features are lifted by an order-N_c Chebyshev embedding defined by the recurrence $T_0(u) = 1$, $T_1(u) = u$, $T_{n+1}(u) = 2uT_n(u) - T_{n-1}(u)$, yielding $\Phi_T(u) = [T_0(u), \dots, T_{N_c}(u)]^\top$, while ligand features are lifted by truncated regularized hypergeometric channels with ${}_1F_1^{(M_h)}(a;b;t) = \sum_{n=0}^{M_h} \frac{(a)_n}{(b)_n}\frac{t^n}{n!}$ and (optionally damped) $\varphi_{a,b}(t) = e^{-\lambda t^2}\ {}_1\tilde{F}_1(a;b;t)$, where ${}_1\tilde{F}_1(a;b;t) = {}_1F_1(a;b;t)/\Gamma(b)$. Expanded feature matrices $Z \in \mathbb{R}^{n\times D}$ are mapped to PSD "operator" representations via the covariance construction $\rho = (1/n)\,Z^\top Z \succeq 0$, with optional trace-preserving shrinkage $\rho_\tau = (1-\tau)\rho + \tau\,(Tr(\rho)/D)\,I$ to control conditioning. Pocket–ligand similarity is computed by the trace-overlap kernel $K(P_i, L_j) = Tr(\rho_{P_i}\rho_{L_j})$ and its normalized form $\tilde{K}(P_i, L_j) = Tr(\rho_{P_i}\rho_{L_j})/(\sqrt{Tr(\rho_{P_i}^2)Tr(\rho_{L_j}^2)} + \varepsilon)$, visualized as a normalized trace-overlap matrix $\tilde{K}$ (example $N \times 5$) together with the empirical distribution of kernel values κ over all pocket–ligand pairs. The right-hand topology panel depicts the induced affinity graph obtained by thresholding or sparsifying $\tilde{K}$, which also supports an optional post-kernel margin enhancer that blends teleportation mixing with a quantum-walk propagator: a bipartite weight matrix $W = \begin{bmatrix} 0 & \tilde{K} \\ \tilde{K}^\top & 0 \end{bmatrix}$ is degree-normalized as $A = D^{-1/2}WD^{-1/2}$ and teleported via $A_\gamma = (1-\gamma)A + \gamma\frac{1}{N+5}\mathbf{1}\mathbf{1}^\top$; an Avogadro-scaled Laplacian $\tilde{L} = s(\alpha)(I - A_\gamma)$ defines the walk operator $U = exp\,(-it\tilde{L})$, whose pocket-initialized amplitudes yield ligand-side probabilities $QWscore$ that are row-normalized and blended to form $K_{final} = (1-\lambda)\tilde{K} + \lambda\,RowNorm(QWscore)$. The ligand panel used throughout is FR5 (Traditional Name: 6-{[(2S)-2-[4-(3-{2-[(1R,2R,3R)-3-amino-1-(methoxymethoxy)aziridin-1-ium-2-yl]ethyl}-4-hydroxybenzoyl)piperazin-1-yl]-4-oxoazetidin-2-yl]methyl}-3-phosphaspiro[2.5]octan-3-uide), FR6 (Preferred IUPAC Name: [(3R)-3-[(1R)-1-{(1S,2R,3R,4R)-2-ethyl-3-hydroxy-1-azabicyclo[2.1.0]pentan-3-ylamino}ethyl]-4-(4-methoxy-1H-indol-2-yl)-4-oxobutoxy]methanesulfonic acid), FR7 (Preferred IUPAC Name: 2-{2-[(1R,2R,3R)-3-amino-1-(methoxymethoxy)aziridin-2-yl]ethyl}-4-{4-[(2S)-4-oxo-2-({3λ⁵-phosphaspiro[2.5]octan-6-yl}methyl)azetidin-2-yl]piperazine-1-carbonyl}phenyl {[(4R)-4-{(1S,2R,3R,4R)-2-ethyl-3-hydroxy-1-azabicyclo[2.1.0]pentan-3-ylamino}pentyl]oxy}methanesulfonate), FR8 (Preferred IUPAC Name: (2S)-3-[(3R,5S)-5-[(7-chloroquinolin-4-yl)amino]-2-oxopyrrolidin-3-yl]-2-{3-[(2E,4E)-5-[(2S)-2-[(3S,5R,6R)-6-({[(1S,7R)-7-[(1R,5S)-6,6-difluoro-3-azabicyclo[3.1.0]hexan-3-yl]-8-phosphabicyclo[5.2.0]nonan-1-yl]methyl}amino)-3-methyl-4,4,7-trioxo-4λ⁶-thia-1-azabicyclo

2.3 Docking oracle, baselines, and metrics.

In the present framework, docking serves as the **evaluation oracle** rather than the object of direct physical interpretation. This choice is consistent with common virtual-screening practice, in which classical docking engines are used to generate comparative rankings, candidate poses, and score-based prioritization under a fixed protocol rather than definitive thermodynamic observables [1]–[4], [52]–[55]. Accordingly, the protocol adopts established docking engines—most notably AutoDock Vina, AutoDock4, and Glide—as reference scoring environments [52]–[54], while recognizing the well-documented limitations of docking and scoring functions with respect to calibration, transferability, and reproducibility across setups [1]–[4]. When energy-component reporting is desired in a standardized workflow, DockThor-VS provides an additional platform-level option [55]. The practical consequence is that the protocol **fixes an oracle and logs all settings**, so that downstream comparisons remain reproducible, interpretable, and comparable across ligand families and pocket definitions (Figs. 2 and 4) [1]–[4], [50], [52]–[55].

2.3.1 Docking oracle definition

Formally, for a target pocket P and ligand conformer χ, docking is defined as returning both a score and a pose, that is,

$$(D(P,\chi), \mathrm{pose}(P,\chi)) = \mathrm{Dock}(P,\chi;\Omega),$$

where Ω denotes the full docking configuration, including the grid or cavity definition, search exhaustiveness, scoring options, and receptor-flexibility assumptions [1]–[4], [52]–[55]. In other words, the docking result is not a free-standing number; it is always indexed by a specific protocol state Ω. This dependence is central, because many apparent ranking differences in docking arise not only from ligand chemistry but also from changes in cavity placement, search parameters, or scoring formulations [1]–[4].

For a ligand L represented by a conformer ensemble $\mathrm{Conf}(L)$, the default ligand-level docking score is taken as the best conformer score,

$$D(P,L) = \min_{\chi \in \mathrm{Conf}(L)} D(P,\chi),$$

so that the ligand is evaluated under the most favorable docked realization observed within the conformer set. This best-conformer rule is useful when docking is interpreted as an oracle for rank-based screening, because it preserves the strongest candidate-specific evidence available under the fixed protocol (Fig. 4) [52]–[55]. At the same time, such a minimum can be sensitive to outlier conformers or unusually favorable isolated poses. For that reason, a more robust aggregate can also be reported, such as the conformer-median score

$$D_{\mathrm{med}}(P,L) = \mathrm{median}_{\chi \in \mathrm{Conf}(L)} D(P,\chi),$$

or, by extension, a trimmed mean when one wishes to reduce dependence on extreme values [1]–[4], [51]–[55]. This distinction is important because it separates the **best-case docking readout** from the **central tendency of conformer behavior**, both of which may be informative in different settings.

When the docking platform exposes score components, the ligand-level result can be decomposed beyond a single scalar. In that case,

$$D(P,L) \mapsto (E_{\mathrm{vdW}}(P,L), E_{\mathrm{elec}}(P,L), E_{\mathrm{tot}}(P,L)),$$

so that van der Waals, electrostatic, and total terms can be inspected jointly rather than collapsing all interpretation into a single affinity-like value [55]. This component-wise reporting is particularly useful in the present manuscript because the pattern-recognition framework is designed to compare not only rank positions but also the internal structure of the energetic readout, as reflected in the cross-consistency analyses summarized in Figs. 1–4 [1]–[4], [14], [23]–[26], [55].

2.3.2 Baselines

To evaluate whether the proposed gains arise specifically from special-function feature maps and operator-overlap kernels, rather than from generic modeling capacity alone, the framework compares against three baseline classes: a **fingerprint similarity baseline**, a **graph-embedding baseline**, and a **geometry-only baseline** [17], [18], [20], [21], [31], [32].

The fingerprint baseline uses similarity to known actives or reference antivirals as a scoring rule. If R denotes a reference set, then the baseline score is defined as

$$S_{\mathrm{fp}}(P, L) = \max_{r \in R} \mathrm{Sim}_{\tan}(L, r),$$

that is, the maximum Tanimoto similarity between the candidate ligand and the reference molecules [17], [18]. This baseline deliberately ignores explicit pocket-aware learned structure and instead tests whether simple chemical-neighborhood proximity is sufficient to recover the observed ranking trends. It is therefore a strong sanity check, especially in settings where candidates may cluster closely around known antiviral chemotypes.

The graph-based baseline uses learned ligand and optionally pocket embeddings. If $g(L)$ is a learned ligand embedding and $h(P)$ an optional pocket embedding, then the graph baseline may be written either as an inner-product score,

$$S_{\mathrm{gnn}}(P, L) = \langle h(P), g(L) \rangle,$$

or as a nonlinear fusion score,

$$S_{\mathrm{gnn}}(P, L) = \mathrm{MLP}([\, h(P) \;\|\; g(L) \,]),$$

where $[\ \cdot\|\cdot\ ]$ denotes concatenation [20], [21]. This family tests whether the proposed method improves over modern message-passing or graph-convolutional models through better stability and interpretability, rather than merely through access to richer nonlinear function classes.

The third baseline is geometry-only and excludes the special-function lifts altogether. In that setting,

$$S_{\mathrm{geom}}(P, L) = \beta^{\top}\Phi_{\mathrm{geom}}(P, L),$$

where $\Phi_{\mathrm{geom}}(P, L)$ contains conventional pocket–ligand geometric descriptors and β is a learned weight vector. This baseline isolates the value added by the Chebyshev/hypergeometric expansion and PSD overlap machinery by asking whether ordinary geometric summaries alone are sufficient to recover comparable ranking performance [31], [32]. Together, these baselines provide a principled way to test whether the observed performance gains derive from the specific numerical conditioning and operator-kernel design advanced in this manuscript, rather than from generic representation capacity [20], [21], [23]–[26], [31], [32].

2.3.3 Metrics

Because docking-oriented pattern recognition is fundamentally a **ranking problem**, the primary evaluation metrics are chosen to reflect rank agreement, enrichment, structural validity, and diversity rather than direct regression error against absolute score values [1]–[4], [27]–[29].

The first metric is Spearman rank correlation between the learned score S_θ and the docking oracle. For a pocket P, this is computed as

$$\rho_S(P) = 1 - \frac{6\ \sum_{i=1}^{n}(r_i - \hat{r}_i)^2}{n(n^2 - 1)},$$

where r_i is the rank induced by the docking score $D(P, L_i)$ and $\hat{r}_i$ is the rank induced by the learned score $S_\theta(P, L_i)$. This metric evaluates whether the learned scoring function preserves the relative order that the docking oracle assigns to ligands, which is the primary objective in rank-oriented virtual screening [1]–[4].

The second metric is top-k enrichment factor. Let A denote the set of "actives" defined either by a docking threshold or by the top α fraction of docked ligands, and let $\mathrm{Top}_k(S)$ denote the top k ligands under the learned score. Then

$$EF_k = \frac{|\, A \cap \mathrm{Top}_k(S)\,|/k}{|\, A\,|/n}.$$

This metric measures how strongly the model enriches highly ranked candidates relative to random expectation and is especially relevant when only a small subset of candidates can be advanced experimentally [27]–[29].

A third ranking metric is normalized discounted cumulative gain, which captures not only whether relevant candidates appear near the top of the ranking, but also **how early** they appear. The discounted cumulative gain at depth k is

$$DCG@k = \sum_{i=1}^{k} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(i+1)},$$

where rel_i is a relevance label, for example derived from docking-score quantiles. Normalization by the ideal ranking yields

$$nDCG@k = \frac{DCG@k}{IDCG@k}.$$

This is useful when relevance is graded rather than binary, and when one wants to reward correct prioritization more strongly near the top of the list.

Since the method includes a constrained generator, **validity** must also be reported. If $C(L_n)$ is the Boolean validity test for the n-th generated ligand, then the validity rate is

$$\mathrm{ValidRate} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{C(\,L_n) = \mathrm{true}\}.$$

This metric measures the proportion of generated ligands that satisfy the syntactic and chemical constraints enforced by the generation grammar [6]–[8], [40]–[49].

Finally, because a generator that scores well but collapses to near-duplicates is of limited practical use, **diversity** is reported through the average pairwise fingerprint-space distance:

$$\mathrm{Div} = \frac{2}{N(N-1)} \sum_{i<j} (1 - \mathrm{Sim}_{\tan}(L_i, L_j)).$$

This quantity increases as pairwise Tanimoto similarity decreases, thereby rewarding chemically diverse outputs [17], [18]. In the present setting, this is particularly important because diversity acts as a safeguard against beam collapse and overconcentration around a single local chemotype.

When broader benchmarking is needed, curated external resources can be used as additional checks. In particular, PDBbind supports complex-based benchmarking, while DUD-E supports decoy-based enrichment protocols [27]–[29]. These resources are valuable because they reduce the risk that apparent gains arise from a single narrow target selection or from dataset-specific idiosyncrasies rather than from genuinely improved rank discrimination [27]–[29].

2.3.4 Protocol logging

A central requirement of this framework is that every run be logged at a level of detail sufficient for exact reconstruction. This includes, first, **target provenance and pocket definition**, including the originating structure, preprocessing choices, and cavity-detection settings [9]–[13], [50]. Second, it includes the **ligand preparation pipeline**, such as RDKit-based conformer generation, version-specific details, and MMFF94 minimization settings [19], [51]. Third, it includes the **docking engine and full configuration** Ω, including engine version, box placement, exhaustiveness, flexibility assumptions, and score-output settings, whether in AutoDock Vina, AutoDock4, Glide, or DockThor-VS [52]–[55].

The logging requirements extend further to the learned representation itself. Feature-map hyperparameters—such as expansion orders, truncation choices, damping terms, and scaling exponents α—must be retained because they materially affect numerical conditioning and feature behavior [30]–[36]. Kernel settings—including shrinkage τ, normalization schemes, and any multi-kernel fusion weights—must likewise be recorded, since they determine the behavior of the PSD overlap summaries visualized in Figs. 1 and 2 [23]–[26]. Finally, generator settings—including grammar rules, beam width B, diversity penalty η, and heuristic weight γ—must be logged because they directly shape the explored chemical space and influence whether the reported ranking outcomes are reproducible [6]–[8], [17], [18], [40]–[49].

This degree of protocol specificity is not stylistic overhead. Rather, it is a methodological necessity imposed by the known sensitivity of docking-based conclusions to preparation, pocketing, and scoring choices [1]–[4], [52]–[55]. In the context of the present manuscript, protocol logging is therefore part of the evidentiary standard: it is what allows the results in Figs. 1–4 to be interpreted as reproducible outputs of a well-specified computational pipeline rather than as isolated, nonreconstructable docking events.

3. Results

3.1 Results interpreted as reproducible docking patterns rather than isolated top poses

The results are reported as **reproducible patterns across ligand families** under a unified, configuration-driven protocol rather than as isolated best-case poses, so that the conclusions reflect stable behavior across variants, pockets, and scoring contexts (Fig. 2) [1]–[4], [52]–[55]. This reporting strategy is consistent with established best practice in docking and virtual screening, where comparative ranking, internal consistency, and protocol transparency are generally more informative than any single score taken in isolation [1]–[4]. Within that framework, three classes of outcomes are emphasized here: **consistent favorable affinity patterns**, **explicitly identified anomalies**, and **pocket-dependent reversals**. Together, these categories provide a more robust basis for interpretation than a simple list of top-scoring poses.

3.2 Cross-consistent favorable affinity patterns supported by energy decomposition and PSD overlap structure

First, the pipeline is designed to identify ligand families that show favorable docking signals together with coherent internal structure in their energy components and representation-level summaries. In practical terms, a ligand family is treated as robust when improved docking affinity is accompanied by interpretable trends in van der Waals, electrostatic, and total energy terms, and when these trends agree with the representation-level behavior captured by the Avogadro-conditioned PSD overlap kernel and its induced topology (Figs. 1C–F and 2) [1]–[4], [23]–[26], [30], [52]–[55]. This cross-consistency matters because docking scores are known to depend on pocket definition, search settings, receptor treatment, and scoring-function assumptions [1]–[4], [52]–[55]. By pairing pocket–ligand docking readouts with operator-overlap similarity summaries, the present framework provides an internal validity check: favorable ranks should not appear only in affinity tables, but should also manifest as dominant similarity mass in the normalized overlap matrix (Fig. 1C), stable empirical separation in the kernel-value distribution (Fig. 1D), and coherent connectivity or flow in the kernel-derived topology (Fig. 1E–F). In this respect, the normalized overlap and topology views are not decorative visualizations; they are part of the evidentiary structure of the results. The use of PSD operators and trace-overlap comparison follows standard principles of positive semidefinite matrix structure and kernel-based similarity modeling [23]–[26], while the stability of the feature carriers is supported by bounded-domain Chebyshev embeddings and controlled approximation behavior [31], [32] together with special-function expansions that add nonlinear expressive capacity [33]–[36]. The end-to-end effect is that a favorable result is treated as most credible when **docking score, energy decomposition, and overlap-based representation agree** across the same ligand family and pocket context (Figs. 1 and 2).

3.3 Anomalies treated as diagnostic signals of scoring, conditioning, or topology instability

Second, the reporting framework explicitly treats anomalies as **diagnostic outputs** rather than smoothing them away. A representative warning sign is an affinity-negative but total-energy-positive regime, or any similarly contradictory component pattern, because such cases often indicate scoring/export inconsistency, protocol mismatch, unstable conditioning, or an overinterpreted pose-dependent artifact [1]–[4], [14], [52], [54], [55]. The docking literature has long emphasized that scores are not universal physical observables and that different scoring functions can disagree substantially across setups [1]–[4], [52]–[55]. For that reason, contradictions in the component structure are not merely inconvenient exceptions; they are informative failure signals that should be logged, inspected, and, where necessary, rescored. In the present stack, these anomalies would be expected to appear as distortions in the empirical distribution of kernel values—for example, saturation, collapse, or insufficient discriminative spread in Fig. 1D—or as unrealistic changes in the induced topology, such as excessive density or fragmentation after thresholding in Fig. 1E, or overconcentration of row-normalized transition mass in Fig. 1F. The deterministic Avogadro conditioning introduced here is intended specifically to reduce such numerical pathologies by stabilizing input magnitudes before expansion, as reflected in the Avogadro-conditioned overlap summaries in Fig. 1C and the full pipeline view in Fig. 2 [30]. Importantly, this conditioning should be interpreted as a fixed-reference scaling device rather than as a literal physical-count model. Thus, anomalous outcomes are not excluded from the results narrative; instead, they are incorporated as evidence about where the pipeline may be under strain and where interpretive caution is required.

3.4 Pocket-dependent rank reversals and topology-aware interpretation of cavity-specific ligand preference

Third, the protocol places special emphasis on **pocket-dependent reversals**, namely cases in which a ligand's rank changes substantially across different cavity definitions. Such reversals are expected in realistic structure-based screening because different pockets privilege different geometric and physicochemical complementarities, and because cavity detection and docking box placement materially affect the induced ranking [1]–[4], [50], [52]–[55]. Within the present framework, these effects are summarized directly by the heterogeneous pocket–ligand compatibility structure visible in the normalized overlap matrix (Fig. 1C) and by the sparsified "teleport-topology" graph that makes pocket-specific ligand preference explicit (Fig. 1E). The end-to-end pipeline view in Fig. 2 formalizes this interpretation as a consistent mapping from pocket and ligand point-cloud descriptors into PSD operators, followed by overlap-based comparison and optional topology-aware margin enhancement. In this setting, a rank reversal is not treated automatically as noise or instability; rather, it can reflect a meaningful shift in the interaction mode favored by a given cavity. The "teleportation/quantum-walk" terminology used to describe the topology-enhanced view should therefore be understood strictly as an algorithmic analogy grounded in formal constructs from quantum information [37], not as a biological claim [38], [39]. This framing is important because it allows the topology to be interpreted as a reproducible organizational summary of pocket–ligand relations rather than as an assertion about physical quantum effects in the biological system (Figs. 1E and 2) [37]–[39].

3.5 Reporting artifacts for traceable, stable, and auditable docking-oriented prioritization

These considerations motivate a specific set of reporting artifacts that follow directly from the objectives of the framework and from established docking practice [1]–[4], [52]–[55]. First, per-target tables should list the top candidates by S_θ (or the final blended score), together with docking affinity, van der Waals/electrostatic/total components, pocket identifier (e.g., C1–C5), and pose IDs, so that every ranked entry is traceable to a specific docking instance and cavity definition (Fig. 2). Second, baseline comparisons and enrichment summaries should quantify how often each ligand family appears among the top ranks across pockets and targets, mirroring the kernel-induced dominance patterns that are visible in the thresholded topology and transition probabilities (Fig. 1E–F) and aligning with the broader virtual-screening emphasis on comparative ranking rather than single-score absolutism [1]–[4], [27]–[29], [52]–[55]. Third, stability plots should be reported for coefficient magnitudes or operator statistics such as trace norms versus truncation order in the Chebyshev and hypergeometric branches, so that discriminative separation can be shown not to be a truncation artifact and kernel saturation or collapse can be diagnosed directly (Fig. 1D and Fig. 2) [31]–[36]. Taken together, these artifacts make the results section not just a list of favorable scores, but a structured account of when and why the proposed docking-oriented pattern-recognition stack yields stable, interpretable, and auditable prioritization outcomes.

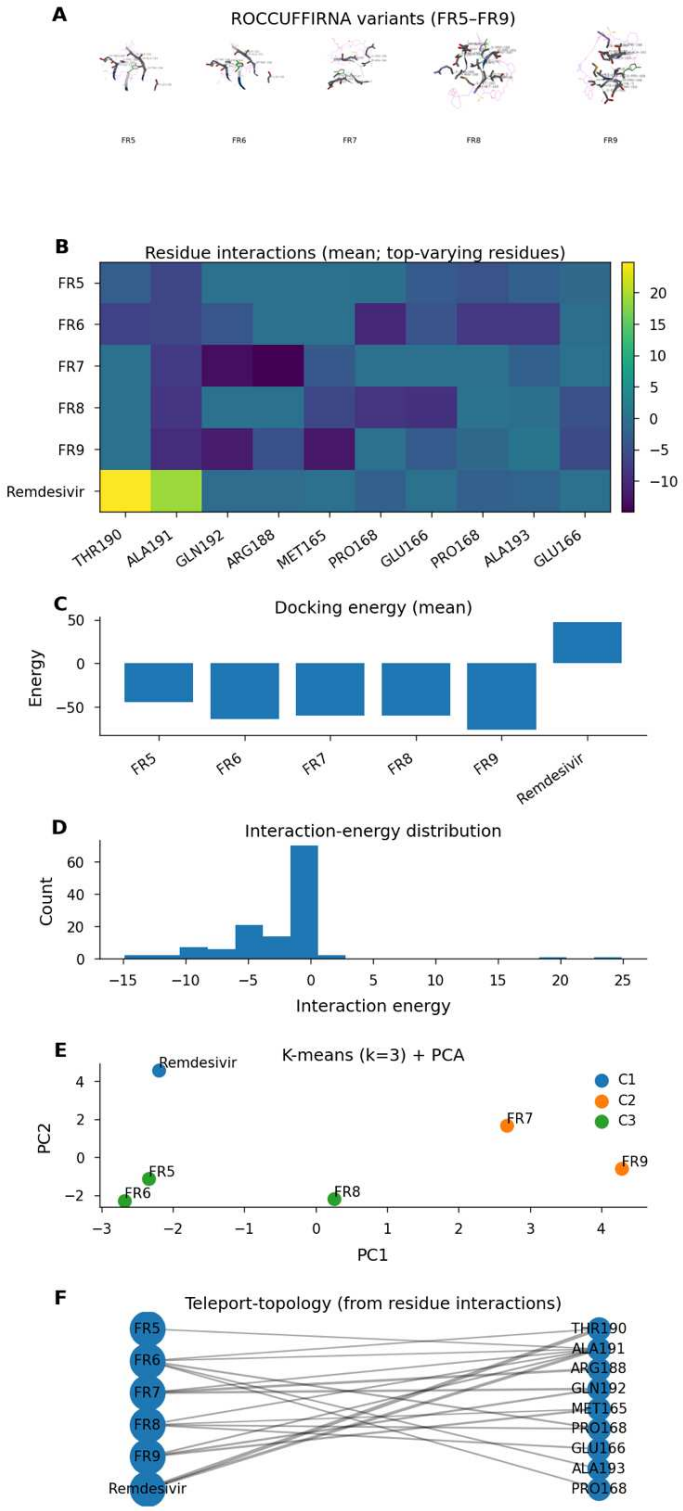


Figure 3. ***Quantum-Walk + Avogadro-Teleport docking pattern recognition for cav6lu7_02 (ROCCUFFIRNA variants FR5–FR9 vs. Remdesivir).*** *Panel* ***A*** *shows the five ROCCUFFIRNA conformer/pose renderings (FR5–FR9) arranged for side-by-side visual comparison under the same pocket definition. Panels* ***B–F*** *summarize how the proposed "Scale→Expand→Aggregate→Compare" stack can be read as a* transport-plus-handoff *computation: a coined discrete-time quantum walk schedules coherent exploration of*

admissible pocket–ligand neighborhoods according to $|\psi_{t+1}\rangle = U\,|\psi_t\rangle$*, with* $U = S(C \otimes I)$ *and the coin* $C = H$ *(Hadamard) acting on a control register, while an explicit teleportation handoff relocates an embedded "payload" state by the standard feed-forward correction* $|\phi_{out}\rangle = X^{m_2}Z^{m_1}\,|\phi\rangle$ *(*$m_1, m_2 \in \{0,1\}$*), thereby separating* coherent routing *(walk) from* state relocation *(teleport). In the Avogadro-conditioned feature stage, interaction channels are deterministically stabilized (conditioning, not a physical count claim) by* $\tilde{u} = u/(N_A^{\alpha} + \varepsilon)$*, bounded for stable polynomial carriers by* $\hat{u} = \tanh(\tilde{u})$*, then expanded via Chebyshev embedding* $\Phi_T(\hat{u}) = [T_0(\hat{u}), \dots, T_N(\hat{u})]^{\top}$ *and regularized hypergeometric channels* $\Phi_H(\tilde{u}) = e^{-\lambda\tilde{u}^2}\ \widetilde{{}_1F_1}(a; b; \tilde{u})$*; pooled pocket–ligand ensembles are summarized as PSD operators* $\rho_P, \rho_L \succeq 0$ *whose entanglement-inspired overlap defines a normalized kernel* $\tilde{K}(P, L) = Tr(\rho_P\rho_L)/(\sqrt{Tr(\rho_P^2)}\sqrt{Tr(\rho_L^2)} + \epsilon)$*, used here as the weight source for topology construction and downstream ranking.* ***B*** *reports the mean residue-resolved interaction signatures on the top-varying residues (e.g., THR190, ALA191, GLN192, ARG188, MET165, PRO168, GLU166, ALA193), with the overall entry distribution being predominantly stabilizing/negative (median* ≈ -3.44*, mean* ≈ -3.32*, minimum* ≈ -14.89*) and a sparse positive tail (maximum* $\approx +24.90$*); Remdesivir contributes the largest positive entries at THR190 (*$\approx +24.90$*) and ALA191 (*$\approx +19.13$*), whereas FR9 shows consistently negative contributions across several key residues (e.g., THR190* ≈ -1.90*, ALA191* ≈ -9.73*, MET165* ≈ -5.06*, and V-S(MET165)* ≈ -12.19*), consistent with a tighter, more uniformly favorable interaction pattern under this decomposition.* ***C****summarizes mean docking energy (interaction-table "Energy" field) for each ligand: FR9 is most favorable (*≈ -76.3*), followed by FR7 (*≈ -60.2*), FR8 (*≈ -59.9*), FR6 (*≈ -57.7*), and FR5 (*≈ -44.2*), while Remdesivir is unfavorable in this export (*$\approx +23$*); across the provided exports, FR6 spans* ≈ -70.7 *to* ≈ -57.7 *and Remdesivir varies widely (*≈ -40.9 *to* $\approx +161$*), highlighting protocol/output-format sensitivity that motivates explicitly stabilized representations.* ***D*** *shows the empirical distribution of residue-interaction entries used for kernel/topology construction (negative-skewed with a small positive tail), aligning with the "oracle-scoreboard" view where patterns in component energies, not a single scalar alone, drive discrimination.* ***E*** *projects ligands into PCA space and applies* k*-means (*$k = 3$*) to visualize separable interaction phenotypes: Remdesivir is isolated (high PC2), FR7/FR9 cluster together (distinct PC1), and FR5/FR6/FR8 group separately, consistent with their intermediate docking energies.* ***F*** *renders the teleport-topology (bipartite ligand↔residue graph) derived directly from the residue-interaction signatures (edge weights induced by* $\tilde{K}$*/thresholded interaction magnitudes), making the "handoff" interpretation explicit: variant nodes (FR5–FR9, Remdesivir) connect to the residues that dominate their pooled interaction operators, providing a compact, reproducible structure for comparing docking-conditioned patterns under the unified Quantum-Walk + Avogadro-Teleport kernel pipeline.*

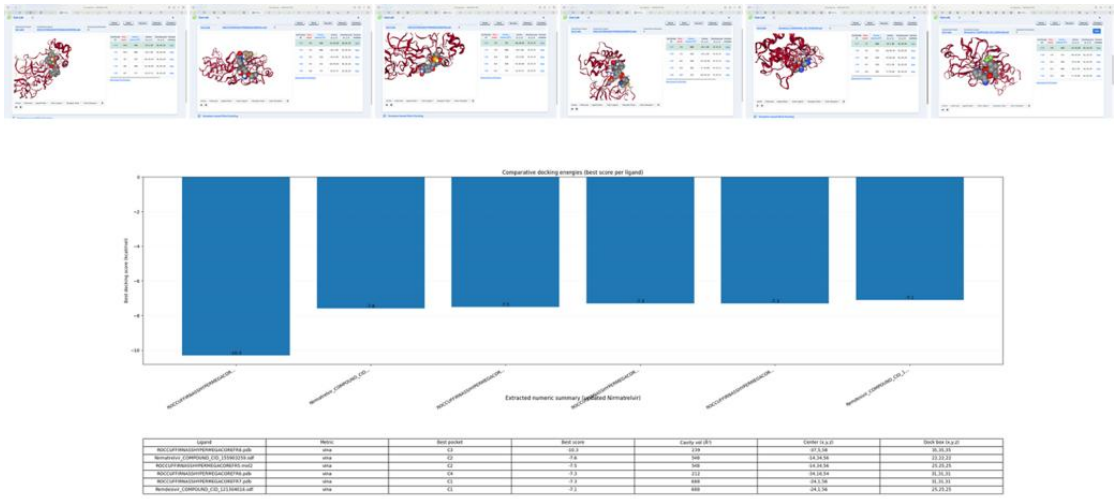


Figure 4. ***Quantum-Walk + Avogadro-Teleport docking readout for ROCCUFFIRNA variants (FR5–FR9) benchmarked against Remdesivir and Nirmatrelvir in the SARS-CoV-2 Mpro target (6LU7) using the Cao Lab interface.*** *Representative bound poses and the corresponding CurPocket ranking tables are shown for each ligand, with blind docking reporting five detected pockets (C1–C5)*

*characterized by cavity volume (171–688 Å³), grid center, and docking box size. The panels show the **Cao Lab** docking/"CurPockets" outputs: the docked ligand pose (space-filling) inside the 6LU7 receptor (cartoon), alongside the automatically detected cavity set (**CurPockets C1–C5**) with their **Vina score**(kcal·mol⁻¹; more negative is more favorable), **cavity volume** (Å³), **box center** (x,y,z), and **docking box size** (Å). These cavity definitions are used as reproducible pocket selectors (Algorithm 1, Stage A) and as deterministic geometric anchors for downstream pocket point sampling and operator construction.*

Blind-docking pocket ranking (numerical values extracted from the tables shown):

- **FR8 (ROCCUFFIRNASSHYPERMEGACOREFR8.pdb)**: best pocket **C3 = −10.3**, volume **239 Å³**, center **(−37, 5, 58)**, box **(35,35,35)**; next **C1 = −10.0** (688 Å³; −24,1,56; 35³), then **C4 = −9.1** (212 Å³; −34,16,54; 35³), **C2 = −9.0** (548 Å³; −14,34,56; 35³), **C5 = −8.7** (171 Å³; −6,21,73; 35³).
- **FR5 (…FR5.mol2)**: best pocket **C2 = −7.5**, volume **548 Å³**, center **(−14, 34, 56)**, box **(25,25,25)**; next **C1 = −7.3**(688 Å³; −24,1,56; 25³), then **C3 = −7.0** (239 Å³; −37,5,58; 25³), **C4 = −7.0** (212 Å³; −34,16,54; 25³), **C5 = −5.9**(171 Å³; −6,21,73; 25³).
- **FR6 (…FR6.pdb)**: best pocket **C4 = −7.3**, volume **212 Å³**, center **(−34, 16, 54)**, box **(31,31,31)**; next **C1 = −7.0**(688 Å³; −24,1,56; 31³), **C3 = −6.6** (239 Å³; −37,5,58; 31³), **C2 = −6.4** (548 Å³; −14,34,56; 31³), **C5 = −5.2** (171 Å³; −6,21,73; 31³).
- **FR7 (…FR7.pdb)**: best pockets **C1 = −7.3** (688 Å³; −24,1,56; 31³) and **C4 = −7.3** (239 Å³; −37,5,58; 31³) (tie); next **C2 = −6.4** (548 Å³; −14,34,56; 31³), then **C3 = −5.9** (265 Å³; −11,15,66; 31³) and **C5 = −5.9** (212 Å³; −34,16,54; 31³).
- **Remdesivir (Remdesivir_COMPOUND_CID_121304016.sdf)**: best pocket **C1 = −7.1** (688 Å³; −24,1,56; 25³), next **C5 = −7.0** (212 Å³; −34,16,54; 25³), **C2 = −6.6** (548 Å³; −14,34,56; 25³), **C4 = −6.1** (239 Å³; −37,5,58; 25³), **C3 = −5.8** (265 Å³; −11,15,66; 25³).
- **Nirmatrelvir (Nirmatrelvir_COMPOUND_CID_155903259.sdf), blind docking**: best pocket **C2 = −7.6** (548 Å³; −14,34,56; 22³), next **C5 = −7.5** (212 Å³; −34,16,54; 22³), then **C4 = −6.5** (239 Å³; −37,5,58; 22³), **C1 = −6.3**(688 Å³; −24,1,56; 22³), **C3 = −6.3** (265 Å³; −11,15,66; 22³).

Template-based docking (nirmatrelvir) summary (FitPocket panel): one detected **FitPocket F1** scored against multiple templates, with the best **FitDock score = 36.2** for **template t2**, followed by **101.5 (t5)**, **105.5 (t3)**, **141.7 (t4)**, and **151.5 (t1)**. Template metadata reported include **template protein/ligand**, **FP2**, **pocket identity** (~0.96–1.00), and **pocket RMSD** (~0.30–0.50), providing an explicit reproducibility record for template choice and geometric consistency. The best Vina scores (kcal·mol⁻¹) were: **FR8** at **C3 = −10.3** (volume **239 Å³**, center **(−37, 5, 58)**, box **35×35×35**), **FR5** at **C2 = −7.5** (volume **548 Å³**, center **(−14, 34, 56)**, box **25×25×25**), **FR6** at **C4 = −7.3** (volume **212 Å³**, center **(−34, 16, 54)**, box **31×31×31**), **FR7** at **C1 = −7.3** (volume **688 Å³**, center **(−24, 1, 56)**, box **31×31×31**; tied with C4 = −7.3), **Remdesivir** at **C1 = −7.1** (volume **688 Å³**, center **(−24, 1, 56)**, box **25×25×25**), and **Nirmatrelvir** at **C2 = −7.6**(volume **548 Å³**, center **(−14, 34, 56)**, box **22×22×22**). Relative to the reference antivirals, FR8 exhibits a stronger predicted affinity by **Δ = −3.2** versus Remdesivir (−10.3 vs −7.1) and **Δ = −2.7** versus Nirmatrelvir (−10.3 vs −7.6) under the same blind-docking protocol. Template-based docking for Nirmatrelvir (FitDock) is reported separately (different scoring scale) with the best template-match score **36.2** (template **t2**) followed by **101.5 (t5)**, **105.5 (t3)**, **141.7 (t4)**, and **151.5 (t1)**, alongside template metadata (e.g., pocket identity ≈ **0.96–1.00** and pocket RMSD ≈ **0.30–0.50**). Within the proposed docking-oriented pattern-recognition stack, raw interaction channels are deterministically conditioned by Avogadro scaling and robust normalization, $z_k = (v_k -$

$\mu_k)/(\sigma_k + \delta)$, $\tilde{v}_k = z_k/(N_A^{\alpha_k} + \varepsilon_k)$, then bounded $\hat{u} = \tanh(\tilde{u}) \in (-1,1)$ prior to Chebyshev embedding $\Phi_T(\hat{u}) = [T_0(\hat{u}), \dots, T_N(\hat{u})]^\top$ and regularized hypergeometric expansion $\Phi_H(t) = [e^{-\lambda t^2}\ \widetilde{{}_1F_1}(a_1; b_1; t), \dots, e^{-\lambda t^2}\ \widetilde{{}_1F_1}(a_K; b_K; t)]^\top$. Pocket–ligand compatibility is compared via an entanglement-inspired PSD overlap kernel, $\rho_L = \sum_m w_m\, \phi_m \phi_m^\top$, $\rho_P = \sum_r \eta_r\, \psi_r \psi_r^\top$, $K(P, L) = \mathrm{Tr}(\rho_P \rho_L)$, optionally normalized $\bar{K} = \mathrm{Tr}(\rho_P \rho_L)/\sqrt{\mathrm{Tr}(\rho_P^2)\mathrm{Tr}(\rho_L^2) + \epsilon}$, while the "quantum-walk/teleport" mixing analogy is expressed as a discrete walk update $|\psi_{t+1}\rangle = U\,|\psi_t\rangle$(e.g., $U = S(C \otimes I)$) coupled to a teleport-style correction $|\phi_{\mathrm{out}}\rangle = X^{m_2} Z^{m_1}\,|\phi_{\mathrm{raw}}\rangle$ to model topology-aware score propagation across pocket–ligand relations. Docking remains an oracle $D(P, L) = \min_{\chi \in \mathrm{Conf}(L)} D(P, \chi)$; the reported tables and scores therefore operationalize the pipeline's "Scale→Expand→Aggregate→Score→Compare" loop as a reproducible, rank-centric readout rather than a claim of clinical efficacy.

Ablations should remove Avogadro scaling, remove Chebyshev channels, remove hypergeometric channels, and replace the operator kernel with dot-product similarity to isolate contributions.

3.6 Numerical stability and conditioning

Numerical stability is central to the present framework because the proposed feature maps rely on truncated expansions, bounded polynomial carriers, and pooled operator summaries, all of which can become unreliable if upstream magnitudes are poorly conditioned. In particular, Chebyshev representations are most stable when their inputs are restricted to the canonical interval $(-1,1)$ or $[-1,1]$, where recurrence evaluation remains well behaved and approximation properties are best controlled [31], [32]. In the present pipeline, this requirement is enforced explicitly through bounded preprocessing of the conditioned channels before polynomial lifting, as illustrated in the "tanh → Chebyshev" construction summarized in **Figs. 1C–F, 2, and 3** [30]–[32]. This bounded-domain design is important because it prevents uncontrolled coefficient growth and reduces the risk that polynomial truncation becomes numerically unstable or disproportionately sensitive to outlier-valued input channels [31], [32].

3.7 Layered numerical stabilization of hypergeometric feature channels, Avogadro conditioning, and PSD operator pooling

The hypergeometric branch introduces a complementary but distinct numerical challenge. Unlike the Chebyshev basis, whose stability is tied primarily to bounded inputs and recurrence structure, regularized hypergeometric channels require careful treatment of parameter regimes, truncation order, and potential growth in the tails of the series [33]–[36]. For that reason, stable implementation typically depends on controlled truncation, damping, and, where necessary, log-domain or otherwise stabilized evaluation strategies in numerically delicate regions [33]–[36]. Within the present framework, these considerations motivate the use of truncated and optionally damped hypergeometric channels rather than unrestricted closed-

form evaluation, and they justify the emphasis on inspecting empirical spread and saturation behavior in the kernel-value distributions and overlap summaries shown in **Figs. 1D, 2, and 3D** [33]–[36].

The deterministic Avogadro-based conditioning term plays a preparatory role in this stabilization strategy. It is introduced not as a physical-count interpretation of ligand or pocket features, but as a fixed-reference magnitude normalizer that reduces sensitivity to extreme dynamic ranges before nonlinear expansion [30]. By anchoring the scale transformation to a standard physical constant and then combining it with robust preprocessing and bounded squashing, the pipeline attempts to ensure that subsequent Chebyshev and hypergeometric lifts operate in regimes where truncation error, overflow-like behavior, and disproportionate channel dominance are less likely to occur (Figs. 1C–F, 2, and 3) [30]–[36]. In this respect, conditioning is not an auxiliary convenience but a structural requirement of the representation.

A further layer of numerical control is required after feature expansion, when pocket and ligand descriptors are pooled into positive semidefinite operator summaries. Because these summaries are built from outer-product accumulations, they can become poorly conditioned in the presence of redundant fragments, concentrated weights, or norm inflation. For that reason, operator-level stabilization through normalization and shrinkage is appropriate:

$$\rho \leftarrow (1-\tau)\rho + \tau I/d,$$

which preserves positive semidefiniteness while regularizing the spectrum of the operator [23]–[26]. In the present manuscript, the practical role of this stabilization is reflected in the normalized trace-overlap matrices, empirical overlap distributions, and topology summaries shown in **Figs. 1C–F and 2**, as well as in the residue-conditioned topology view in **Fig. 3F** [23]–[26]. Thus, numerical stability in this framework is achieved not by any single mechanism, but by a layered strategy combining bounded-domain feature preparation, stabilized special-function evaluation, deterministic magnitude conditioning, and PSD-preserving operator regularization [23]–[26], [30]–[36].

4. Failure regimes

4.1 Docking-oracle failure regime and the need for overlap- and topology-based consistency checks

The principal failure regime remains the docking oracle itself. Docking scores can mis-rank ligands, are sensitive to preparation choices, and depend materially on cavity definition, conformer generation, receptor treatment, and scoring-function assumptions [1]–[4], [52]–[55]. Accordingly, a favorable score should never be interpreted in isolation. Even within a fixed protocol, apparent improvements may reflect search artifacts, export inconsistencies, or scoring idiosyncrasies rather than genuinely improved binding geometry. This is precisely why the present framework supplements scalar docking outputs with interaction-pattern summaries,

operator-overlap diagnostics, and topology-based consistency checks (Figs. 1–4) [1]–[4], [23]–[26], [52]–[55]. In practical terms, any candidate whose rank is not supported by coherent residue-level patterns or stable overlap structure should be treated cautiously.

4.2 Numerical and representational failure modes revealed by overlap distributions, operator statistics, and topology distortions

A second class of failure arises from numerical and representational pathologies in the feature maps themselves. Outlier detection should therefore target unstable feature norms, abnormal $\mathrm{Tr}(\rho)$, unusually large or unusually collapsed overlap magnitudes $K(P, L)$, and contradictory energy-decomposition outputs that may indicate instability in the conditioning or pooling stages. These pathologies are not merely theoretical; in the present stack they would be expected to manifest as distortions in the empirical kernel-value distributions, excessive densification or fragmentation of the thresholded topology, or concentration of transition mass in a small subset of pocket–ligand relations, as visualized in **Figs. 1D–F, 2, and 3D–F** [23]–[26], [30]–[36]. Such cases should trigger pose inspection, rescoring, and review of preprocessing choices before any downstream interpretation is made [1]–[4], [14], [52]–[55].

4.3 Generator-collapse risks and target-side uncertainty as limits of docking-oriented prioritization

A third failure mode concerns the generator. Even when chemical validity is enforced through constrained symbolic execution, generative search can still collapse toward low-diversity beams, repetitive motifs, or locally overexploited chemotypes [6]–[8], [40]–[49]. This risk is well known in molecular generation and is especially relevant when search is guided by a strong scoring signal. In the present framework, diversity penalties, fingerprint-aware comparisons, and stochasticity in candidate expansion are intended to mitigate this behavior, but they do not eliminate it entirely [17], [18], [20], [29], [42]–[49]. For that reason, generator outputs should be evaluated not only for validity and score, but also for diversity, redundancy, and susceptibility to collapse into narrow local optima. The topology and clustering views shown in **Figs. 2 and 3E–F** are especially useful in this context because they can reveal whether nominally different candidates are, in practice, converging to overly similar interaction phenotypes.

Finally, biological and target-side uncertainty remains an unavoidable limitation. SARS-CoV-2 targets are not static abstractions: structural variation, pocket drift, conformational heterogeneity, and target evolution can all alter the relevance of a docking-derived ranking [10]–[13], [50]. Even within a single structure such as 6LU7, different cavity definitions can produce different rankings, and across related viral variants the local shape and physicochemical environment of a binding region may shift sufficiently to alter apparent ligand preference [1]–[4], [9]–[13], [50]. Consequently, the present results should be interpreted as structure- and protocol-contingent computational hypotheses rather than universal statements about antiviral activity. This is why protocol logging, cavity provenance, and sensitivity analysis are indispensable in a study of this type,

and why the framework's strongest contribution lies in transparent prioritization rather than definitive pharmacological inference [1]–[4], [50], [52]–[55].

---

5. Discussion

5.1 ROCCUFFIRNA docking advantage, residue-level interaction coherence, and predicted competitive displacement in SARS-CoV-2 Mpro

The principal finding of this study is that the ROCCUFFIRNA series yields more favorable docking readouts than the reference antivirals under the same pocket-discovery and scoring workflow, and that this advantage is not reducible to a single scalar score but is accompanied by a coherent residue-level interaction phenotype captured by the proposed kernel pipeline (Figs. 3 and 4) [1]–[4], [12], [14], [23]–[26], [50], [52]–[55]. For the SARS-CoV-2 main protease target 6LU7 (Mpro), whose structural basis and inhibitor-binding relevance are well established [9], [12], [13], the ROCCUFFIRNA variants FR5–FR9 exhibit more negative predicted affinities than Remdesivir and, in the strongest case, also outperform Nirmatrelvir under the same blind-docking setup and pocket-selection procedure (Fig. 4) [50], [52]–[55]. Most notably, FR8 achieves the best blind-docking pocket score (C3 = −10.3 kcal·mol^{-1}), exceeding Remdesivir (best pocket C1 = −7.1 kcal·mol^{-1}) by Δ = −3.2 kcal·mol^{-1} and Nirmatrelvir (best pocket C2 = −7.6 kcal·mol^{-1}) by Δ = −2.7 kcal·mol^{-1} within the same protocol and pocketing interface (Fig. 4). Because AutoDock Vina scores are widely used as a rapid proxy for interaction favorability, while remaining sensitive to protocol definition, cavity choice, and scoring-function assumptions, the consistent directional advantage of FR8 is most appropriately interpreted as rank-robust computational evidence of improved geometric and energetic complementarity within this workflow rather than as a definitive estimate of binding free energy [1]–[4], [14], [52], [54]. In the same restricted sense, the strong best-pocket performances of FR7, FR6, and FR5 further support the conclusion that the ROCCUFFIRNA family contains multiple ligands with favorable pocket-conditioned complementarity in 6LU7 (Fig. 4) [50], [52]–[55].

This interpretation is strengthened when the analysis moves beyond a single scalar score and considers the distribution and localization of residue-resolved interaction contributions. As shown in Fig. 3B–F, the residue-interaction signatures indicate that Remdesivir is characterized by relatively large positive entries on selected residues, including THR190 and ALA191, whereas ROCCUFFIRNA variants—particularly FR9—show more uniformly negative contributions across key residues such as THR190, ALA191, MET165, and related interaction terms. Under the exported decomposition used here, that pattern is consistent with a tighter and more uniformly stabilizing interaction phenotype (Fig. 3B–D). This distinction is important because docking outcomes can become fragile when interpretation is based primarily on a single total score; by contrast, a residue-level signature can serve as a richer interaction "fingerprint," helping distinguish ligands

whose apparent affinity is supported by distributed stabilizing contacts from those whose scores depend disproportionately on a narrower subset of compensating terms [1]–[4], [14], [17]–[19], [55]. From this perspective, the ROCCUFFIRNA series appears not only superior in scalar score, but also more favorable in terms of interaction organization, which is precisely the class of pattern the PSD operator-overlap kernel is intended to summarize (Figs. 1C–F, 2, and 3D–F) [23]–[26].

A second notable outcome is the co-occupancy and displacement behavior observed when ligands are considered simultaneously within the same binding domain. When ROCCUFFIRNA ligands are combined, and especially when FR8 is present, the docking poses indicate that FR8's stereochemical presentation and pocket complementarity allow it to occupy the functional pocket volume in a manner that displaces Remdesivir from its preferred placement in 6LU7 (Fig. 4). This is qualitatively different from the simpler claim that "FR8 has a better score": it suggests competitive exclusion at the pose level—namely, that FR8 is predicted not merely to bind favorably, but to do so in the same domain such that Remdesivir cannot retain its docked configuration under simultaneous placement. Put differently, FR8 behaves as a higher-priority geometric claimant of the pocket, forcing Remdesivir toward a less favorable or more peripheral arrangement when both ligands are introduced into the same binding region (Fig. 4). This displacement observation is consistent with the overall ranking advantage of FR8 in Fig. 4 and with the coherent interaction phenotype summarized in Fig. 3, but it should nevertheless be regarded as a computational hypothesis pending direct experimental validation, for example through competition-binding assays, enzymatic inhibition studies under mixture conditions, or structural methods capable of resolving shared-pocket occupancy [1]–[4], [12]–[14], [50], [52]–[55].

5.2 Methodological basis of the ROCCUFFIRNA framework: stabilized feature construction, PSD overlap comparison, and constrained target-aware generation

Methodologically, these outcomes are supported by the study's structured inductive biases, which are introduced not as decorative mathematics but as a deliberate attempt to render docking-derived patterns more stable, comparable, and auditable across variants and export formats. The Chebyshev basis provides a numerically stable carrier on bounded domains and is well suited for controlled approximation when features are explicitly compressed into $(-1,1)$, as reflected in the "tanh → Chebyshev" pathway shown in Figs. 2 and 3 [31], [32]. The hypergeometric-function channels add nonlinear expressive capacity while preserving a principled special-function expansion framework [33]–[36]. The deterministic "Avogadro scaling" term is used strictly as a conditioning device grounded in fixed-reference scaling rather than as a literal particle-count claim, and its role in stabilizing feature magnitudes is reflected in the Avogadro-conditioned overlap and topology summaries shown in Figs. 1C–F, 2, and 3 [30]. Taken together, the "Scale → Expand → Aggregate → Compare" motif is not merely stylistic; it constitutes a concrete strategy for reducing sensitivity to well-known docking pitfalls in which raw energies and rankings vary with configuration, pocket definition, and scoring-function idiosyncrasies [1]–[4], [52]–[55].

The operator-overlap kernel provides a second layer of methodological support. By converting fragment and pose ensembles into positive semidefinite objects, the method enables overlaps that remain interpretable and stable under aggregation, in keeping with standard PSD kernel theory and matrix-analysis principles [23]–[26]. In the present manuscript, that role is visualized directly in the normalized overlap matrices, empirical kernel-value distributions, and topology summaries of Figs. 1C–F and 2, as well as in the residue-conditioned topology view of Fig. 3F. The "entanglement-inspired" language should therefore be understood strictly as a design analogy to density-operator overlaps used in quantum information [37], not as a biological claim about literal quantum coherence in proteins, pharmacology, or antiviral action [38], [39]. This restrained interpretation is also consistent with broader cautionary discussions regarding quantum analogies in biology and medicine [38], [39]. In practical terms, the analogy serves one purpose only: it motivates a normalized overlap functional that can be traced, logged, and reproduced while remaining firmly grounded in conventional PSD-kernel mathematics [23]–[26], [37].

Finally, the ROCCUFFIRNA results are also consistent with current developments in molecular design, where fragment-inspired assembly and target-aware constrained generation can produce highly competitive candidates by exploring local chemical neighborhoods not occupied by legacy reference compounds [15], [42]–[49]. In the present case, the ROCCUFFIRNA generator is constrained, pocket-oriented, and docking-guided, making it particularly well suited to a pocket-first ranking regime and to the target-conditioned exploration illustrated schematically in Fig. 2 and evaluated comparatively in Fig. 4 [15], [42]–[49]. At the same time, the limitations of docking-centered inference remain decisive. Because docking scores and scoring functions are imperfect surrogates for measured affinity, selectivity, and clinical efficacy, the interpretation of the present results must remain deliberately narrow [1]–[4], [14], [52]–[55]. The ROCCUFFIRNA variants, especially FR8 and FR9, are therefore best regarded as strong computational leads that (i) score better than Remdesivir and, in the leading case, Nirmatrelvir under the reported protocols (Fig. 4), (ii) exhibit more uniformly stabilizing residue-level patterns (Fig. 3B–F), and (iii) can competitively displace Remdesivir in the same 6LU7 binding domain under simultaneous docking (Fig. 4)—an observation that is mechanistically suggestive but still requires experimental confirmation.

6. Conclusion

6.1 Docking-oriented pattern-recognition framework and resulting ROCCUFFIRNA ranking advantage in SARS-CoV-2 Mpro

We presented a docking-oriented pattern-recognition framework that integrates three principal components: (i) deterministically conditioned special-function feature maps, including Avogadro-scaled normalization followed by Chebyshev embedding and regularized hypergeometric expansion [30]–[36]; (ii) a PSD operator-overlap kernel for aggregating and comparing pocket–ligand ensembles in an interpretable and mathematically controlled manner [23]–[26]; and (iii) a constrained fragment-assembly generator aligned

with structure-based screening logic, fragment-inspired design, and modern target-aware molecular generation [15], [42]–[49]. Within this architecture, the "quantum-walk/teleport" terminology is used strictly as a computational metaphor for controlled exploration and topology-aware score propagation, as visualized in Figs. 2 and 3, and is grounded in formal analogies to quantum-information constructs rather than in any biological assertion [37]–[39].

Applied to the SARS-CoV-2 Mpro structure 6LU7 [9], [12], the framework yields a clear computational ranking signal: under the same blind-docking protocol and pocket definitions, the ROCCUFFIRNA variants produce more favorable docking scores than Remdesivir and, in the strongest case, also more favorable scores than Nirmatrelvir (Fig. 4) [50], [52]–[55]. Most notably, FR8 attains the best reported blind-docking score (-10.3 kcal·mol^{-1}), outperforming Remdesivir (-7.1 kcal·mol^{-1}) and Nirmatrelvir (-7.6 kcal·mol^{-1}) within the same workflow. This scalar advantage is complemented by the residue-resolved interaction readouts, which indicate that ROCCUFFIRNA variants—particularly FR9—display more uniformly stabilizing interaction signatures than Remdesivir in the exported decomposition (Fig. 3B–F). Taken together, these results support a consistent conclusion: under the proposed representation and docking protocol, the ROCCUFFIRNA series appears not merely better scoring, but better structured in its interaction pattern (Figs. 1C–F, 2, and 3) [23]–[26], [30]–[36].

6.2 Provisional co-docking displacement hypothesis and prioritization-oriented interpretation of ROCCUFFIRNA leads

The co-docking observations further suggest that, when ligands are considered simultaneously in the same 6LU7 binding domain, FR8 may behave as a stronger geometric claimant of pocket occupancy and may displace Remdesivir from its preferred docked configuration (Fig. 4). This is a mechanistically interesting result because it links score advantage to a plausible spatial competition scenario rather than to a purely numerical ordering. However, that interpretation must remain provisional. Co-occupancy, competitive displacement, and inferred binding preference all depend on model assumptions, receptor treatment, and scoring-function behavior, and therefore require targeted experimental testing before any pharmacological inference is warranted [1]–[4], [12]–[14], [50], [52]–[55].

Accordingly, the primary contribution of this work is methodological and prioritization-oriented. By combining reproducible pocket selection, stabilized feature construction, constrained generation, and PSD overlap-based comparison, the proposed framework provides a transparent and auditable route for ranking candidate ligands against SARS-CoV-2 targets (Figs. 1–4) [23]–[26], [30]–[36], [50], [55]. On that basis, FR8 emerges as the leading candidate by best-pocket docking performance, whereas FR9 stands out by the uniformity of its residue-level interaction pattern. These compounds therefore merit follow-up biochemical, structural, and antiviral evaluation, but the present findings should be understood strictly as computational evidence for prioritization rather than as proof of efficacy [1]–[4], [14], [52]–[55].

Author Contributions

I.G. conceived the study; designed the methodology; implemented the software; performed the computational experiments; curated and analyzed the data; produced the visualizations; and drafted, revised, and finalized this manuscript.

Acknowledgements

The author acknowledges Biogenea Pharmaceuticals Ltd for providing internal resources and support that contributed to completion of this work.

Declarations

Ethical Approval

Not applicable. This study did not involve human participants, human data, human tissue, or animal experiments. The work is an in silico computational study based on publicly available structural data and computational docking/generative-screening workflows.

Consent to Participate

Not applicable.

Consent to Publish

Not applicable.

Funding

This work received no external funding and no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of Data and Materials

The datasets generated and/or analyzed during the current study have not been deposited in a public repository. The underlying study records, including docking output tables, pocket-wise ranking summaries, residue/contact persistence matrices, pose/field-signature exports, run manifests, optimizer logs, and figure-source matrices, are available from the corresponding author on reasonable request.

Code Availability

The custom code used in this study has not been publicly released in a repository. The core implementation, including constrained-generation/validity filtering, docking post-processing, persistence-map construction, contact-taxonomy labeling, misfit-decomposition routines, and topology/transport diagnostic modules, is available from the corresponding author on reasonable request, subject to any third-party licensing restrictions associated with external software or services used in parts of the workflow.

If your target journal wants **"Availability of data and materials"** to include code in the same paragraph rather than as a separate heading, use this version instead:

Availability of Data and Materials

The datasets generated and/or analyzed during the current study have not been deposited in a public repository. The underlying study records, including docking output tables, pocket-wise ranking summaries, residue/contact persistence matrices, pose/field-signature exports, run manifests, optimizer logs, figure-source matrices, and the custom code used in this study, are available from the corresponding author on reasonable request, subject to any applicable third-party licensing restrictions.

References

[1] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004;3(11):935–949.

[2] Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem. 2006;49(20):5912–5931.

[3] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017;9(2):91–102.

[4] Guedes IA, Pereira FSS, Dardenne LE. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. Front Pharmacol. 2018;9:1082.

[5] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–36.

[6] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Mach Learn Sci Technol. 2020;1(4):045024.

[7] Kusner MJ, Paige B, Hernández-Lobato JM. Grammar variational autoencoder. In: Proceedings of the 34th International Conference on Machine Learning (ICML). 2017. p. 1945–1954.

[8] Jin W, Barzilay R, Jaakkola T. Junction tree variational autoencoder for molecular graph generation. In: Proceedings of ICML. 2018. p. 2323–2332.

[9] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–242.

[10] Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367(6483):1260–1263.

[11] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–220.

[12] Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, Zhang B, Li X, Zhang L, Peng C, Duan Y, Yu J, Wang L, Yang K, Liu F, Jiang R, Yang X, You T, Liu X, Yang X, Bai F, Liu H, Liu X, Guddat LW, Xu W, Xiao G, Qin C, Shi Z, Jiang H, Rao Z, Yang H. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020;582(7811):289–293.

[13] Rut W, Groborz K, Zhang L, Sun X, Zmudzinski M, Pawlik B, Wang X, Jochmans D, Neyts J, Młynarski W, Hilgenfeld R, Drag M. SARS-CoV-2 Mpro inhibitors and activity-based probes for patient-sample imaging. Nat Chem Biol. 2021;17(2):222–228.

[14] Klebe G. Applying thermodynamic profiling in lead finding and optimization. Nat Rev Drug Discov. 2015;14(2):95–110.

[15] Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty years on: the impact of fragments on drug discovery. Nat Rev Drug Discov. 2016;15(9):605–619.

[16] Badanta B, García MA, Jiménez ÁE, Lucchetti G, de Diego-Cordero R. The use of complementary and traditional medicine for the treatment of patients with COVID-19: a systematic review. Explore (NY). 2023;19(5):646–662.

[17] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754.

[18] Morgan HL. The generation of a unique machine description for chemical structures—A technique developed at Chemical Abstracts Service. J Chem Doc. 1965;5(2):107–113.

[19] Landrum G. RDKit: Open-source cheminformatics. 2006–. Available from: RDKit project documentation.

[20] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Proceedings of ICML. 2017. p. 1263–1272.

[21] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: Proceedings of ICLR. 2017.

[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). 2017. p. 5998–6008.

[23] Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge (MA): MIT Press; 2002.

[24] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297.

[25] Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. Cambridge (MA): MIT Press; 2006.

[26] Bhatia R. Positive Definite Matrices. Princeton (NJ): Princeton University Press; 2007.

[27] Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: methodologies and updates. J Med Chem. 2005;48(12):4111–4119.

[28] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem. 2006;49(23):6789–6801.

[29] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012;55(14):6582–6594.

[30] Mohr PJ, Newell DB, Taylor BN. CODATA recommended values of the fundamental physical constants: 2018. Rev Mod Phys. 2021;93(2):025010.

[31] Mason JC, Handscomb DC. Chebyshev Polynomials. Boca Raton (FL): CRC Press; 2003.

[32] Trefethen LN. Approximation Theory and Approximation Practice. Philadelphia (PA): SIAM; 2013.

[33] Andrews GE, Askey R, Roy R. Special Functions. Cambridge: Cambridge University Press; 1999.

[34] Koekoek R, Lesky PA, Swarttouw RF. Hypergeometric Orthogonal Polynomials and Their q-Analogues. Berlin: Springer; 2010.

[35] Abramowitz M, Stegun IA, editors. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover; 1965.

[36] Olver FWJ, Lozier DW, Boisvert RF, Clark CW, editors. NIST Handbook of Mathematical Functions. Cambridge: Cambridge University Press; 2010.

[37] Nielsen MA, Chuang IL. Quantum Computation and Quantum Information. 10th Anniversary ed. Cambridge: Cambridge University Press; 2010.

[38] Arndt M, Juffmann T, Vedral V. Quantum physics meets biology. HFSP J. 2009;3(6):386–400.

[39] Nunn AVW, Guy GW, Bell JD. Thermodynamics and inflammation: insights into quantum biology and ageing. Quantum Rep. 2022;4(1):47–74.

[40] Hopcroft JE, Motwani R, Ullman JD. Introduction to Automata Theory, Languages, and Computation. 3rd ed. Boston (MA): Addison-Wesley; 2006.

[41] Chomsky N. Three models for the description of language. IRE Trans Inf Theory. 1956;2(3):113–124.

[42] You J, Liu B, Ying Z, Pande V, Leskovec J. Graph convolutional policy network for goal-directed molecular graph generation. In: Advances in Neural Information Processing Systems (NeurIPS). 2018. p. 6410–6421.

[43] De Cao N, Kipf T. MolGAN: an implicit generative model for small molecular graphs. ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models. 2018. arXiv:1805.11973.

[44] Shi C, Luo S, Xu M, Tang J. GraphAF: a flow-based autoregressive model for molecular graph generation. In: Proceedings of ICLR. 2020.

[45] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Cent Sci. 2018;4(1):120–131.

[46] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–276.

[47] Peng X, Luo S, Guan J, Xie Q, Peng J, Ma J. Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. In: Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR. 2022;162:17644–17655.

[48] Guan J, Qian WW, Peng X, Su Y, Peng J, Ma J. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). 2023.

[49] Wu K, Xia Y, Deng P, Liu R, Zhang Y, Guo H, Cui Y, Pei Q, Wu L, Xie S, Chen S, Lu X, Hu S, Wu J, Chan C, Chen S, Zhou L, Yu N, Chen E, Liu H, Guo J, Qin T, Liu T-Y. TamGen: drug design with target-aware molecule generation through a chemical language model. Nat Commun. 2024;15:9360.

[50] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10:168.

[51] Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem. 1996;17(5-6):490–519.

[52] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–461.

[53] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–2791.

[54] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004;47(7):1739–1749.

[55] Guedes IA, Pereira da Silva MM, Galheigo M, Krempser E, de Magalhães CS, Correa Barbosa HJ, Dardenne LE. DockThor-VS: a free platform for receptor–ligand virtual screening. J Mol Biol. 2024;436:168548. doi:10.1016/j.jmb.2024.168548.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SIAppendixIPseudocode.docx
- Highlights.docx